ANALYSIS AND SYNTHESIS OF SPEECH BASED ON AN HUMAN AUDITORY MODEL

BY

MINKYU LEE

A DISSERTATION PRESENTED TO THE GRADUATE SCHOOL
OF THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

UNIVERSITY OF FLORIDA

1996

to my parents

ACKNOWLEDGMENTS

I would like to thank my advisor and committee chairman, Dr. Donald G. Childers, for his invaluable guidance and assistance in the Mind–Machine Interaction Research Center (MMIRC) throughout this research. His standards and style were a source of inspiration and encouragement. I am also thankful for his financial assistance.

I would also like to thank Dr. Leon W. Couch, II, Dr. Jose C. Principe, Dr. Fred J. Taylor and Dr. Howard B. Rothman for their invaluable time and interest in serving on my supervisory committee.

I would like to thank all my colleagues in the MMIRC. I have had a great time with my colleagues during this study. All advice and critiques from them should be a great asset through out my professional life.

I am deeply indebted to my parents and parents–in–law for their love and encouragement. Last but not least, I wish to thank Sooha, my wife, and Seungyeon, my daughter, for their love, faith and patience throughout the years.

TABLE OF CONTENTS

APPENDICES

Abstract of Dissertation Presented to the Graduate School
of the University of Florida in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy

ANALYSIS AND SYNTHESIS OF SPEECH BASED ON AN HUMAN AUDITORY MODELING

By

Minkyu Lee

May 1996

Chairman: Dr. D.G. Childers
Major Department: Electrical and Computer Engineering

Most speech analysis and synthesis systems have focused on voiced speech, which has stationary and deterministic characteristics. However, little effort has been devoted to the analysis of unvoiced speech. That is mainly because of the difficulty in analyzing unvoiced speech, which is characterized by its aperiodic characteristics.

In spite of the technical difficulties, research for unvoiced speech is essential for various applications. In speech coding systems, vocoders based on a linear source–filter model have been widely used for low–bit–rate (2400–4800 bits per second) speech coding. In order to reduce the bit rate lower than 2400 bits per second, many ideas have been proposed. Speech coding based on the articulatory production model is one approach and promises good quality at very low bit rates. The position vector of articulators that generates a speech segment is a good candidate for efficient speech coding. For this purpose, there must be a reliable method to estimate the articulatory positions from the acoustic speech signal. This procedure is called speech inverse filtering and should be able to produce a reliable result not only for voiced speech but also for unvoiced speech. In order to get a good inverse filtering result, we need more knowledge about the unvoiced speech production mechanisms. However, an understanding of unvoiced speech production is still in a preliminary stage.

From experiments, it has been determined that unvoiced speech is highly related to the vocal tract front cavity resonance. The noise source for unvoiced speech is located near the vocal tract constriction and the front cavity serves as a spectral shaping filter. The purpose of this research is to study the effect of the front cavity resonance and the vocal tract area function on the quality of synthesized unvoiced speech.

An algorithm based on the human auditory model is used to estimate the front cavity resonance. Using the front cavity resonance frequency estimate, the effective length of the vocal tract front cavity can be calculated. The turbulent noise source spectrum is also estimated using the front cavity resonance information. The parameters obtained from the analysis phase are combined to construct a simple vocal tract area function to generate unvoiced speech. Using an articulatory synthesizer, which consists of a turbulent noise source generator and a vocal tract filter, unvoiced speech is generated, and effects of the front cavity length, constriction length, and back cavity resonance on the perception of unvoiced speech are studied by informal listening tests. Effects of the spectrum of the turbulent noise source on the synthesized speech are also investigated.

CHAPTER 1
INTRODUCTION

In the research area of speech analysis, synthesis, recognition, compression, and coding, a main concern is how speech is generated using the articulators – tongue, teeth, lips, etc. In order to understand the mechanism of speech production, various speech production models have been proposed. These production models are explained in this chapter followed by an overview of speech synthesis techniques. Then, based upon the production models and speech synthesis methods, we provide the details of speech analysis–by–synthesis, which is widely accepted in the research area. After that, motivations of this research will be given.

1.1 Speech Production Model

There are two kinds of speech production models. One is the linear source filter model, which approximates the speech waveform as a linear convolution of the glottal source waveform and the vocal tract transfer function (Fant, 1960). The other model is the articulatory production model that simulates the physiological mechanisms of human speech production. Most speech application methods in use today are based on the linear source tract model because it is simple in its structure, easy to calculate its parameters and the synthesized speech is typically of good quality. One of the obstacles in achieving a practical system based on the articulatory model is the difficult problem of estimating articulatory parameters from the speech signal.

1.1.1 Linear Source Filter Model

In the linear filter model of speech production as depicted in Figure 1–1 (Fant, 1960), the speech signal is modeled as the convolution of a linear quasi–time–invariant

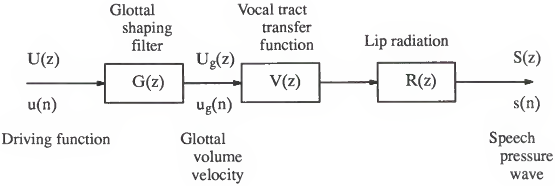


Figure 1–1. Block diagram representation of the linear speech production model.

filter excited by quasi–periodic pulses for voiced sounds, or by random noise for unvoiced sounds, or by both pulses and random noise for mixed excitation sounds. In other words, speech is assumed to be generated when one or more sound sources excite the vocal tract filter, which is followed by a lip radiation filter. Therefore, from the viewpoint of speech analysis based on the linear source–filter model, a challenging problem is to analyze the acoustic speech waveform in order to separate the speech into a vocal tract component and a glottal source component. The lip radiation filter can always be approximated by a simple differentiator regardless of the source type and the vocal tract shape. According to the linear speech production model, the vocal tract component can be thought of as a filter transfer function and the glottal source component as an excitation signal to the filter. The transmission characteristic of the vocal tract is well approximated by a cascade of uncoupled resonators and anti–resonators whose bandwidths and center frequencies may be independently controlled.

A general discrete time model based on the linear source filter model is depicted in Figure 1–2 (Rabiner and Schafer, 1978). Numerous speech production models that are based upon the linear source–filter model all have in common that the excitation are separated from the vocal tract and radiation part. Often, it is also assumed that the excitation source and the vocal tract shape are independent, which means the nonlinear source–tract interaction is not included in this model. In fact, it is this assumption that makes the linear source–filter model so simple and easy to understand. The excitation generator creates a quasiperiodic pulse train for the voiced speech and a noiselike random signal for the unvoiced speech. The time–varying linear system accounts for the vocal tract transfer function and lip radiation effect. Parameters of the model can be chosen so that the output of the model is well matched to the desired real speech according to a predefined error criteria. From previous research, it is accepted that the vocal tract component determines the phonetical information, while the glottal source characteristics are related to the vocal quality, i.e., whether the voice is normal, breathy, or vocal fry

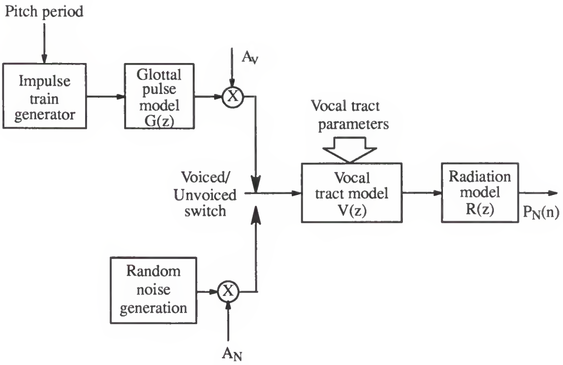


Figure 1–2. General discrete–time model of speech production (After Rabiner and Schafer, 1978).

(Childers and Ahn, 1995; Childers and Lee, 1991; Eskenazi *et al.*, 1990). This result can be directly applied to speech synthesis and speech recognition. Therefore estimating the glottal input waveform from the speech is a major research topic in the speech analysis area.

There are three primary ways to obtain the glottal volume velocity waveform. One method is to use high speed photography to measure a glottal area function. Fant(1986) simulated the impedances of the subglottal and supraglottal tracts, leading to a time–varying relationship between the glottal area and glottal flow waveform. Another method is to use a reflectionless tube to neutralize the effect of the vocal tract transfer function (Sondhi, 1975). The third method is glottal inverse filtering (Klatt and Klatt, 1990). The glottal inverse filtering (GIF) is based on the linear source filter model, which implies that it is possible to extract the glottal source waveform from the speech signal provided a good approximation of the vocal tract transfer function is available.

The fundamental problem in obtaining an accurate estimate of the glottal source waveform by glottal inverse filtering is to determine the parameters of the inverse filter of the vocal tract transfer function. Numerous ways to obtain the characteristics of the vocal tract transfer function from the speech signal have been suggested. However, no method has proven to be universally acceptable for all speech data. For example, the closed phase method has been shown to give good results for AR parameters in cases where the speech has a distinct glottal closed phase. The formant frequencies and bandwidths can be obtained from the roots of the predictor polynomial. But when the speech has a high fundamental frequency or has a short closed phase as in a woman's or a child's voice, the results are not reliable. For vocal disordered speech, the closed glottal interval often does not exist or is so short that the closed phase LP algorithm often fails to give good estimates of the vocal tract parameters.

1.1.2 Articulatory Model

While most speech analysis and synthesis methods use the linear model, which is not well correlated with the mechanism of human speech production, the articulatory model attempts to model the mechanical motion of the articulators and the volume velocity and sound pressure in the lungs, larynx, and vocal and nasal tracts (Flanagan *et al.*, 1975). From the articulatory point of view, the vocal tract can be modeled as an acoustic tube with nonuniform and time-varying cross-sections. There are two major classes of articulatory models: parametric area models and midsagittal distance models. The parametric models regard the vocal tract cross sectional area as a function of distance along the vocal tract from the glottis to the lips, while the midsagittal models specify the positions of the tongue body, tongue tip, jaw, lips, hyoid, and velum in the midsaggital plane. The overall shape of the vocal tract can be converted from the articulatory parameters.

The articulatory models require several orders of magnitude more computation than the linear models and the quality of the speech synthesized with the articulatory models is not comparable with that of the linear formant models. Nonetheless, as the articulatory model is very close to the human speech production mechanism, it is likely to provide the ultimate solution to producing natural and intelligible speech synthesis by computer (Klatt, 1987). In the speech coding/compression area, it is possible to use these articulatory parameters as potential candidates for efficient coding, e.g., very low bit-rate speech coding and/or communication since the parameters vary slowly in the human voice production system (Flanagan *et al.*, 1980).

1.1.2.1 Area models

These models represent the vocal tract as an area function without using knowledge of the articulatory positions directly. They concentrate on modeling the cross sectional area of the vocal tract as a function of distance along the vocal tract subject to some constraints (Fant, 1960; Atal *et al.*, 1978; Flanagan *et al.*, 1980; Lin, 1990; Yu, 1993). Their common feature is a specification of the minimum constrictional area A_c and its axial location X_c. The area of the vocal tract is usually represented by a continuous function such as a hyperbola, a parabola, or a sinusoid (Lin, 1990). Consonant articulations have generally not been implemented. Figure 1–3 shows one example of an area function obtained from a parametric area model.

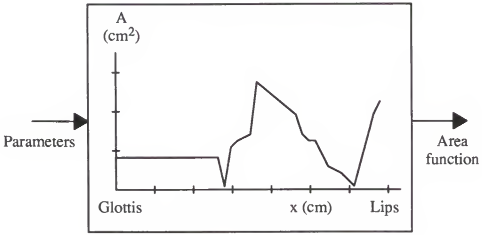


Figure 1–3. A parametric area models.

1.1.2.2 Midsagittal distance models

The midsagittal distance models are usually based on a representation of the midsagittal plane as seen from an X-ray image. Figure 1–4 shows one example of midsagittal distance models, which describes the speech organ movements in a midsagittal plane and requires an input to specify the positions of the articulators (Mermelstein, 1973;

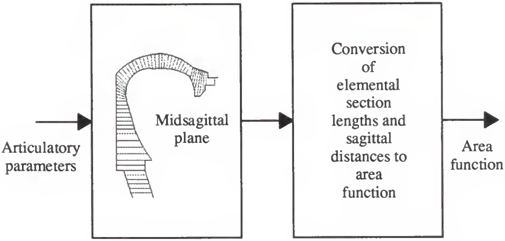

Articulatory parameters
Midsagittal plane
Conversion of elemental section lengths and sagittal distances to area function
Area function

Figure 1–4. An example of midsagittal distance models.

Levinson and Schmidt, 1983; Sondhi and Schroeter, 1986; Prado, 1991) or to control the movements of the articulators by rules (Coker, 1976; Parthasarathy and Coker, 1990, 1992). The output is an estimate of the vocal tract cross-sectional area. Visualization and articulatory state interpretations are the major advantages of these models.

1.2 Speech Analysis and Synthesis

A fundamental goal of speech analysis and synthesis is to acquire a basic understanding of the speech communication processes, which can be applied to the efficient encoding, transmission and processing of speech information. Speech analysis is the process of estimating the parameters of the speech production model from a speech signal that is assumed to be the output of that model. Continuous speech is usually analyzed by performing analysis processes repeatedly on short segments of the speech signal (typically 10–30 msec durations), producing time–varying parameters of the model. One of the goals of speech analysis research is to obtain good estimates of the speech production model parameters. In the source–filter model, the analysis output is a set of parameters defining the shape of the excitation source waveform and the vocal tract transfer characteristics. In the case of the articulatory model, the parameters describe the location and/or movement of the articulators. The accuracy of the estimates are assessed by the quality of synthetic speech using the model parameter estimates.

Speech synthesis is the process of producing an acoustic signal by controlling and updating the speech production model with an appropriate set of parameters. The model parameters can be obtained either by the analysis of real speech signals or by the analysis–by–synthesis procedure. If the model is sufficiently accurate and its parameters are accurately estimated, the resulting output of the model should be comparable to natural speech.

1.2.1 Speech Synthesis

The synthesis of speech has been studied in great detail. The fundamental principles of human speech generation mechanism are well understood except for some of the time-varying and nonlinear characteristics of vocal fold vibration and source-tract interaction. There are a wide range of applications of speech synthesis and it can contribute to a better understanding of human speech production. For instance, speech synthesis provides a model for evaluating the speech parameters obtained by speech analysis. Speech synthesis can also be utilized in the study of phonetics, since each acoustic parameter can be controlled independently and arbitrarily, thus allowing researchers to evaluate the effect of the acoustic parameters on the characteristics of the synthetic speech. There are essentially three different methods used for speech synthesis: formant synthesis, linear prediction (LP) synthesis, and articulatory synthesis.

1.2.1.1 Formant synthesis

At present, the formant synthesizers can be categorized into two groups: the cascade/parallel formant synthesizer developed by Klatt (Klatt, 1980; Klatt, and Klatt, 1990) and the versatile parallel formant synthesizer developed by Rye and Holmes (1982). It is generally agreed that the Klatt's cascade/parallel model has been favoured for text-to-speech synthesis, while the Holmes' parallel model tends to be used for synthesis-by-analysis systems. The reasons for this are probably related more to the way in which the different synthesis models are controlled, rather than the inherent capabilities of the synthesizers themselves. It has been demonstrated that high-quality speech can be generated with such synthesizers (Klatt and Klatt, 1990).

Clearly, the formant synthesizers do not have direct relation to articulatory movement. Instead, these models use a functional approximation to the effects of varying glottal impedance and subglottal coupling, the subtleties of vocal fold motion, etc. In fact, many successful speech-synthesis systems are based on the formant synthesis although it does not use the same mechanisms as human speech production.

1.2.1.2 Linear prediction (LP) synthesis

In 1971, a new technique was developed for analyzing and synthesizing speech using computers. The method, known as linear prediction (Atal and Hanauer, 1971), has been widely accepted in speech analysis and synthesis. The basic idea is that a signal sample, s(n), can be estimated by a linear combination of its past signal samples s(n–k)

$$\hat{s}(n) = \sum_{k=1}^{p} a_k \, s(n-k) \tag{1–1}$$

where $\hat{s}(n)$ is the estimated signal at instant n, and p is the order of the linear predictor. The linear predictive coefficients, a_k, are determined by minimizing the total error, E, which is the sum of the squared differences, e(n), for a sequence of N samples

$$E = \sum_{n=1}^{N} e^2(n) \tag{1–2}$$

where

$$e(n) = s(n) - \hat{s}(n) = s(n) - \sum_{k=1}^{p} a_k \, s(n-k) \, . \tag{1–3}$$

By transforming Eq.(1–3) to the Z–domain and rearranging it, we obtain

$$S(z) = \frac{1}{1 - \sum_{k=1}^{p} a_k Z^{-k}} \, E(z) = V(z) \, E(z) \tag{1–4}$$

where V(z) is an all-pole transfer function

$$V(z) = \frac{1}{1 - \sum_{k=1}^{p} a_k Z^{-k}} \, . \tag{1–5}$$

The linear predictive (LP) speech synthesizer is a mathematical realization of the linear source–tract model. The LP synthesizer assumes that the speech is generated by an impulse or white noise excitation. The synthetic speech produced by this model is usually intelligible, but often exhibits unnatural characteristics. The typical synthetic speech often sounds "buzzy," producing pop and click sounds mainly because of the poor modeling of the glottal excitation source. In addition, nasal and obstruent sounds are difficult to reproduce because the model is an all-pole model (Childers and Wu, 1990). To improve the quality, the error signal (residue between the predicted signal and the original speech) can be used as the excitation. In recent years, multipulse excitation has improved the quality of LP synthetic speech (Atal and Remde, 1982; Singhal and Atal, 1989). More recent research has shown that various LP synthesizers, such as the glottal excited linear predictive (GELP) synthesizer, can reproduce high quality speech (Childers and Hu 1994; Milenkovic, 1993).

By applying the above basic idea to speech signals, the typical block diagram of an LP synthesizer is constructed as shown in Figure 1–5 (Markel and Gray, 1976). Two types of excitation sources are switched at the input of the all-pole system in order to simulate two different modes of phonation, i.e. voiced and unvoiced speech. In generating voiced sounds, the excitation source is a periodic pulse train and controlled by the pitch period and gain parameter. For unvoiced sounds, the excitation source, E(z), is a random noise, which is controlled by the gain parameter and spectral parameter. The transfer function, V(z), is characterized by the LP coefficients and implemented by a slowly time-varying digital filter. Since it is more like a mathematical model, the parameters controlling the process for the LP synthesizer are completely automatic and does not require formant tracking and spectral fitting procedures. Because of its simplicity, the all–pole system is more popular than the pole–zero system in modeling speech signals. Furthermore, this model has proven to be quite satisfactory in practice.

One of the disadvantages of the LP technique is that it can not acoustically decompose the speech signals into an excitation source and a model of the vocal tract. The reason is that the LP minimization process assumes that the excitation error, e(n), is white random noise, while the glottal excitation source for the voiced speech is a periodic pulse–like waveform (Fujisaki and Ljungqvist, 1986). Therefore, the LP scheme is not suitable for the physiological study of the human speech production system.

1.2.1.3 Articulatory synthesis

While both of the formant and LP synthesis methods are acoustic domain models in which no interaction between the glottal-flow excitation function and the vocal-tract

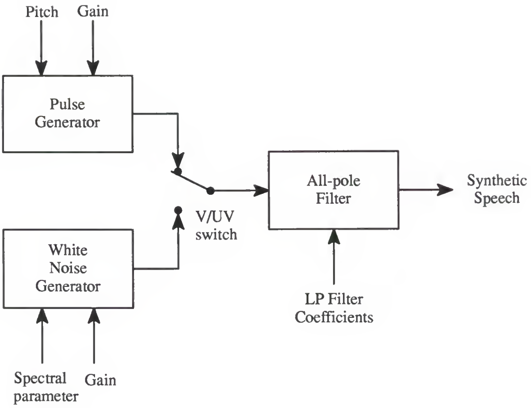


Figure 1–5. Basic structure of LP synthesis.

filter function is involved, the articulatory synthesis is the production of speech sounds using a model of the vocal tract, which simulates the mechanical movements of the speech articulators – glottis, tongue, jaw, teeth, and lips – and the volume velocity, sound pressure in the lungs, larynx, and vocal/nasal tract. Since the articulatory synthesizer is based on a model of the physiology of the human speech production process, nonlinear effects that can not be incorporated in the other synthesis methods are introduced in this model. Not only the coarticulation effects but also the interaction between the vocal tract and the vocal folds, the contribution of the subglottal system and the effects of the nasal tract and sinus cavities are modeled in the articulatory synthesizer.

Articulatory synthesis usually consists of two separate components: an articulatory model and an acoustic model as shown in Figure 1–6. The articulatory model represents the articulatory positions and converts them into vocal tract cross-sectional area functions. The vocal tract is divided into many small sections and the corresponding cross-sectional areas are used as parameters to represent the vocal tract characteristics in terms of the vocal tract cross sectional area. The vocal tract shape can also be drawn from the midsaggital distance model. The acoustic model, which includes subglottal coupling, source-tract interaction, vocal tract, nasal tract with sinus cavities, and acoustic radiation, simulates the speech sound propagation through the vocal system as well as the physics of the physiological-to-acoustic transformation. In the acoustic model, each cross-sectional area is approximated by an electrical analog transmission line to calculate the transfer function of the vocal tract. In order to simulate the movement of the vocal tract, the area functions change with time. Each sound is designated in terms of a target configuration and the movement of the vocal tract is specified by either fast or slow motion of the articulators.

One of the most difficult problems in articulatory synthesis is the lack of analysis data. Analysis procedures usually require an "acoustic-to-articulatory inverse transformation" from the speech signal, i.e., speech inverse filtering. At the present time

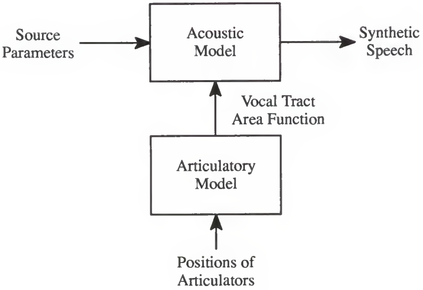


Figure 1–6. Basic structure of articulatory synthesis.

the complexity of the speech inverse filtering is very high due to the non–uniqueness property of the inverse problem. In addition, the optimization algorithm also needs to be improved (Hsieh, 1994). Another way is to use high speed x–rays to measure the articulator positions. This method is also limited by the x–ray dosage for the speaker. The resolution of the film and the speed are other limitations to this approach. The complexity of the relationship between articulatory gestures and the acoustic signal makes it difficult to generate automatically the details of articulatory control needed to produce a synthetic copy of a given sample of human speech (Hsieh, 1994).

Despite these drawbacks, articulatory speech synthesis has several advantages. First, a properly constructed articulatory synthesizer is capable of reproducing all the naturally relevant effects for the generation of fricatives and plosives, modeling coarticulation transitions as well as source-tract interaction in a manner that resembles the physical process that occurs in real speech production. Since the articulatory model has a direct relation to the human speech production process. the articulatory synthesis may lead to natural sounding speech synthesis, which in turn can lead to a simpler and more elegant synthesis–by–rule, e.g., text-to-speech applications.

Secondly, to the extent that we can accurately obtain the speech gestures (articulatory movements or trajectories), articulatory synthesizers may be valuable for research scientists and physicians, since the synthesizers can be used to study linguistic theories, to provide a feedback mechanism for teaching speech production, and to explore the effects of vocal tract surgical techniques on speech production prior to surgical intervention (Childers, 1991); and they hold out the ultimate promise of high quality, natural-sounding speech with a simple control scheme (Klatt, 1987).

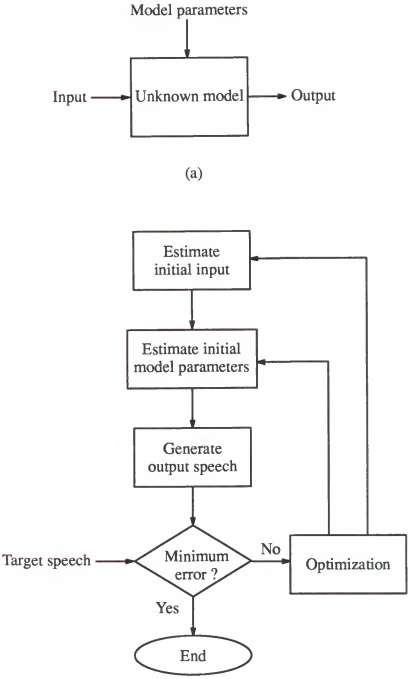
Articulatory synthesizers will continue to be of great importance for research purposes, and to provide insights into various acoustic features of human speech. Thus, an articulatory synthesizer may provide both an efficient description of natural speech and a means for synthesizing natural-sounding speech. However, a major problem with the articulatory synthesizer is the lack of a means to derive articulatory configurations from the speech signal using speech inverse filtering.

There are three major approaches to articulatory synthesis based upon the articulatory production model. The first group solves differential equations in the time domain (Boccieri and Childers, 1984). The second method uses wave digital filters to model the forward–backward traveling waves (Meyer *et al.*, 1989). The third method uses a hybrid time–frequency domain technique (Sondhi and Schroeter, 1987). These articulatory models usually include an excitation source model, for example, Fant (1985) used the volume–velocity waveform model as a glottal excitation source model.

Fant (1960), and Johansson *et al.*, (1983) measured the cross–sectional area of the vocal tract. Sondhi and Resnick (1983), Charpentier (1984), Milenkovic (1987) and Xue *et al.* (1990) developed acoustic–to–articulatory transformations to help derive the vocal tract cross–sectional area from the speech signal directly.

1.2.2 Speech Analysis–by–synthesis

There are several methods for estimating the speech production model parameters from its input and/or output. If both input and output of the model are given, then the problem is simply reduced to a system identification problem in which designing the model and estimating its parameters are the major concern. However, when only the output is known the problem is an optimization problem, which is the case for speech production modeling. For the speech research, the analysis–by–synthesis method often gives good results for estimating the model parameters since the speech production model is usually too complicated to apply an analytic method. The analysis–by–synthesis method for the estimation of the articulatory parameters from the speech signal, which is also called the speech inverse problem, can be illustrated as in Figure 1–7. In the speech



(a)

(b)

Figure 1–7. A basic diagram of analysis–by–synthesis procedure.

inverse problem, the input is the glottal source, which is either a quasi–periodic pulse train for the voiced speech or random noise for unvoiced speech. The model parameters are the position vectors of the articulators, which determine the shape of the vocal tract. In order to find the input and model parameters, initial guesses of the input and articulatory model parameters are used to produce an output speech. The output speech is compared with the target speech to calculate an error and the initial input parameter and model parameters are adjusted to reduce the error. An error between the synthetic speech and target speech can be defined in numerous ways. By repeating this procedure until the error meets some pre–defined threshold value, a set of the articulatory parameters can be obtained.

However, there is no guarantee that the obtained parameters are optimal, since the optimization procedure can be trapped in local minima. Many algorithms have been suggested to solve this problem. Another difficult problem is that the mapping from the articulatory parameter to the output speech is not a one–to–one mapping which means there can be more than one inverse solution for one target speech (Atal *et al.*, 1978). The non–unique property of the speech inverse problem can be solved by using proper initial estimates of the input and model parameters (Sorokin, 1994) The speech inverse problem will be discussed further in the next chapter.

1.3 Motivation

Most speech analysis and synthesis procedure have focused on voiced speech, which is relatively long, quasi–stationary and easy to analyze. Numerous algorithms have been proposed and many of them give successful results for voiced speech. The vocal tract transfer function can be estimated using various spectrum estimation algorithms and the glottal source waveform can be extracted by glottal inverse filtering algorithms. The inverse filter is usually obtained from an estimate of the vocal tract transfer function. It is generally accepted that vocal tract characteristics determine the phonetic information and the glottal source characteristics determine the vocal characteristics, i.e. voice quality such as normal, breathy, whisper, etc.

On the other hand, research on unvoiced speech has not received much attention, mainly because adequate models are not available. In addition, the length of unvoiced speech segments in a sentence are usually short and transient, making it harder to analyze than the voiced speech. Due to these reasons, estimating the articulatory parameters for the unvoiced speech has not been as successful as that for voiced speech.

1.4 Research Goals and Contributions

With this background, the specific goals of this research are as follows:

$\boxed{1}$ To propose a new method for estimation of the front cavity resonance frequency characteristics from unvoiced speech.

$\boxed{2}$ To separate the unvoiced speech spectrum into a turbulence noise source spectrum and a vocal tract transfer function (uncoupled front cavity transfer function).

$\boxed{3}$ To estimate the location of articulation, i.e., location of the constriction in the vocal tract.

The results can lead to an improved articulatory production model that is able to describe unvoiced speech as well as voiced speech. The results of this research can be applied to a low–bit–rate speech coding and natural–sounding speech synthesis.

The contributions of this research will be :

$\boxed{1}$ To propose a new model for the generation of unvoiced fricatives.

$\boxed{2}$ To investigate the effects of turbulence noise source characteristics on the generation of unvoiced fricatives.

$\boxed{3}$ To estimate the location of articulation, i.e., location of the constriction in the vocal tract.

$\boxed{4}$ To develop a solution for estimating the articulatory model parameters for the unvoiced speech, i.e., to solve the inverse problem for unvoiced speech.

1.5 Descriptions of Chapters

The contents of this research report are as follows. Chapter 2 reviews the previous research for the voiced and unvoiced speech analysis including glottal inverse filtering. Next, the psychoacoustical basics are summarized and the perceptual analysis of speech for unvoiced speech is introduced. We also discuss perceptual linear prediction, which is the major tool used to study unvoiced speech generation model. Finally, selective inverse filtering used for manual inverse filtering is discussed.

Chapter 3 describes the design approach of this research, which includes discussions about previous research for unvoiced speech generation models. The research method, analysis procedure, and the limitations of the research are discussed.

Chapter 4 details the analysis procedure and results for the unvoiced source characteristics. Included are the results for the estimation of the vocal tract shape from the analysis of unvoiced fricatives.

Finally, chapter 5 summarizes the research results and concludes with recommendations for future work.

CHAPTER 2
SPEECH INVERSE FILTERING AND PSYCHOACOUSTICAL MODEL

2.1 Speech Inverse Filtering

To analyze a speech signal represented in discrete form is to find a set of parameters of a model that generates the speech signal. For the linear source–filter model, the parameters are usually the coefficients of the autoregressive (AR) model of the vocal tract and the parameters of the glottal volume velocity model. For the articulatory model, the parameters are the position vectors of the articulators. The simple linear source–filter model produces a synthetic speech with acceptable quality. However, it is not capable of modeling the non–linear source–tract interaction nor the nonlinear vocal fold vibration effects that might be critical factors in achieving natural sounding synthetic speech. For this reason, speech research based on the articulatory model is attracting more attention from researchers because of its potential ability to model the sophisticated and nonlinear processes of human speech generation.

According to the articulatory production model, the articulatory parameters can be transformed directly to the vocal tract area function. Given the vocal tract cross–sectional area function, calculation of formant frequencies has been well established in the acoustic theory of speech production (Fant, 1960; Atal et al., 1978; Wakita and Fant, 1978; Badin and Fant, 1984; Fant, 1985; Lin, 1990, 1992). The area function can be mapped to the vocal tract transfer function including the effect of the lip radiation. The overall sequence is depicted in Figure 2–1. The mapping from the articulatory parameters to the area function, i.e., the conversion of elemental section length and sagittal distance to the area

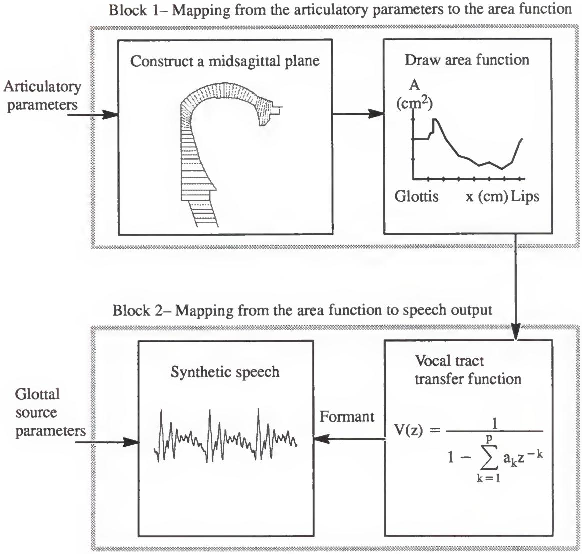

Block 1– Mapping from the articulatory parameters to the area function
Articulatory parameters
Construct a midsagittal plane
Draw area function
A
(cm^2)
Glottis
x (cm) Lips
Block 2– Mapping from the area function to speech output
Glottal source parameters
Synthetic speech
Formant
Vocal tract transfer function
$V(z) = \frac{1}{1 - \sum_{k=1}^{p} a_k z^{-k}}$

Figure 2–1. Generating speech from an articulatory production model using a formant synthesizer.

function (block 1 in Figure 2–1) is not one–to–one mapping. Since the transformation is from a three–dimensional space (articulatory position) to a two–dimensional space (area), there is a possibility that two or more different articulatory configurations may have the same area function. It is known from previous research that many area functions can be transformed into one vocal tract transfer function (Atal *et al.*, 1978). Therefore, in general, transforming the articulatory parameters to a synthetic speech output is a many–to–one mapping, making the inverse problem hard to solve.

Due to the nonunique property of the inverse problem, it is not easy to solve the problem directly from natural speech. Instead, an approach from a different viewpoint is taken. The analysis–by–synthesis method described in the previous chapter has been widely used (Sorokin, 1994). A solution of the inverse problem will help improve not only speech production research, but also assist speech recognition, text–to–speech synthesis, low bit rate speech coding and data compression area.

2.1.1 Applications of the Inverse Problem (Low–bit–rate Coding)

One of the applications that benefit from a successful inverse solution is the speech coding research area. Despite the emergence of inexpensive optical fibers, there is a growing need for bandwidth conservation in the wireless cellular and satellite communication fields. Currently, speech coding research is aimed at the low bit rate (3 ~ 8 kbits/s) and very low bit rate (below 3 kbits/s) coding techniques because the current coding algorithms give unacceptably poor speech quality at low bit rates. For the low bit rate (lower than 16 kbits/s) speech coding, vocoders have been widely used in which the speech waveform is usually represented by a linear source filter model of speech production. Recently, speech coding based on an articulatory production model has received considerable attention since the parameters of the articulatory production model change slowly, and, therefore, promise a highly economic representation of the speech signal. More importantly, due to advances in computing power, it has become possible to analyze speech in order to obtain the articulatory parameters within a reasonable amount of computing time.

The basic idea is that when speech is generated, the articulators change their positions slowly compared with the rate of vibration of the vocal folds. If the positions of the articulators, i.e., the vocal tract cross sectional area, can be estimated from the speech signal, then it can be used as criteria for speech segmentation. One obstacle for achieving a practical system based on this idea is the lack of accurate and efficient methods for estimating the articulatory parameters from the speech signal. If the inverse problem can be solved successfully, then it is possible to establish a relationship between speech segmentation and articulatory movement for the low bit rate speech coding. More detail algorithm is described in Appendix B.

2.2 The Inverse Problem for Voiced Speech

The inverse problem for the voiced speech has been accomplished in two steps. The first step, called glottal inverse filtering, is to decompose the speech signal into the vocal tract transfer function and the glottal excitation source. The vocal tract parameters are further processed to estimate the articulatory parameters. The glottal excitation source waveform can be parameterized using glottal source models. In the second step, the articulatory position parameters are estimated by optimization techniques using the formant information, which is extracted from the vocal tract transfer function. The glottal inverse filtering and glottal source modeling as well as previous research for speech inverse filtering will be reviewed in this section.

2.2.1 Glottal Inverse Filtering

For several decades, researchers have been searching for a solution to the problem of estimating the glottal volume velocity waveform directly from the acoustic speech signal. Some solutions estimate the vocal tract transfer function by the linear prediction algorithm or equivalent spectral estimation methods and do the inverse filtering to obtain the glottal volume velocity waveform (Wong *et al.*, 1979; Veeneman and B$_E$Ment, 1985; Childers and Ahn, 1995). Others estimate the glottal volume velocity waveform as well as the vocal tract transfer function at the same time using iterative methods (Milenkovic, 1986; Fujisaki and Ljungqvist, 1986; Fujisaki and Ljungqvist, 1987; Lobo *et al.*, 1992). Pitch synchronous analysis usually gives more accurate results than pitch asynchronous analysis. Some researchers use EGG signal to find the glottal closed phase and opening phase (Veeneman and B$_E$Ment, 1985; Childers and Ahn, 1995). The vocal tract characteristics are directly related to the position of the articulators, and, therefore, can be used to estimate the articulatory parameter vectors. The glottal volume velocity waveforms are used to study the effect of the glottal wave shape on the vocal quality of the speech. In this section, a brief review of previous results for estimating the glottal volume velocity waveform are given.

One important result for estimating the glottal source waveform and the vocal tract transfer function was the closed phase linear prediction analysis (Wong *et al.*, 1979). By applying the covariance method of linear prediction to the closed glottal interval, good estimates of the vocal tract transfer function were obtained. It was also shown that the glottal closure and opening instants could be determined from the normalized total squared error. Using the estimation of the vocal tract transfer function, the glottal volume velocity could be estimated by inverse filtering. One disadvantage of the closed phase LP algorithm is that the analysis frame length could be too short to obtain a reliable estimation of the vocal tract transfer function since the analysis window is confined only to the glottal closed interval. For a normal male voice with low fundamental frequency this is not a major problem, but for a female or a child's voice, which often has a high fundamental frequency with an incomplete glottal closure interval, the closed phase LP algorithm often fails to provide a reliable estimate of the vocal tract transfer function and the glottal volume velocity waveform (Veeneman and B$_{E}$Ment, 1985).

Fujisaki and Ljungqvist (1987) have noted that the conventional speech analysis method based on linear prediction often fails to separate the source and tract characteristics because the linear prediction scheme assumes the source to be a white noise excitation, while in actuality it is periodic. This assumption causes errors in the estimation of the formant frequencies and bandwidths because the speech spectrum includes not only the vocal tract transfer function but also glottal source information. A polynomial source model was proposed and the source model was combined with the linear predictive analysis in order to separate the source parameters and the vocal tract parameters. In an improved version of this scheme, the approach was extended to apply to a wider variety of speech sounds by introduction of pole–zero modeling for the vocal tract transfer function.

Lee (1988) developed a two–pass algorithm. In the first pass, the speech input is analyzed by a fixed frame length linear prediction algorithm to approximate the vocal tract characteristics. Using these characteristics, the speech is inverse filtered to get the point of glottal excitation that is the starting point for a pitch synchronous analysis frame in the second pass. In the second pass, pitch synchronous LP analysis is performed to get a more accurate vocal tract transfer function. Using this refined transfer function, the differentiated glottal waveform can be obtained by inverse filtering.

Ting (1989) and Lee (1992) used the WRLS–VFF algorithm to get the vocal tract transfer function estimates. It was found that the negative peaks in the variable forgetting factor (VFF) signal indicate the instants of glottal closure, and, therefore, the VFF signal was used to find the starting point of the glottal closed phase region. This algorithm shows good performance even when the glottal closed region is short. However, the algorithm is computationally expensive.

Alku (1992) suggested a new method called pitch synchronous iterative adaptive inverse filtering (PSIAIF) that is based on the iterative adaptive inverse filtering (IAIF). In this algorithm, the glottal contribution to the speech spectrum is first estimated using an iterative procedure. The vocal tract characteristics are obtained by applying linear prediction to the speech data after eliminating the average glottal contribution pitch synchronously. This method can be regarded as an extension of the two–pass algorithm in which more than two iterations are performed.

Milenkovic (1986) suggested a polynomial model for the glottal source waveform for jointly estimating the glottal source waveform parameters and the vocal tract transfer function. He assumed an arbitrary set of glottal source parameters and estimated the vocal tract transfer function using the input source (glottal waveform) and the speech output. Once the input and output are known, the problem becomes a system identification problem. He performed the same procedure repeatedly until all combinations of the glottal source parameters are tried exhaustively. He then chose the best glottal source and vocal tract transfer function pair that generated the minimum error. The EGG signal was used to identify the glottal opening point. The results show very reliable glottal opening instance estimates for both male and female voices.

In summary, the drawback of the closed phase linear prediction algorithm can be avoided by not limiting the analysis frame to the glottal closed interval and using an iterative procedure (Milenkovic, 1986; Alku, 1992) or exhaustive search of the glottal source parameter space for the optimal parameter set (Fujisaki *et al.*, 1986; Fujisaki *et al.*, 1987). In spite of these efforts, however, there are a number of obstacles to overcome to achieve a good glottal inverse filtering algorithm. Most current algorithms work well only for modal male voices. More robust and reliable algorithms that work for female voices and pathological voices must be developed.

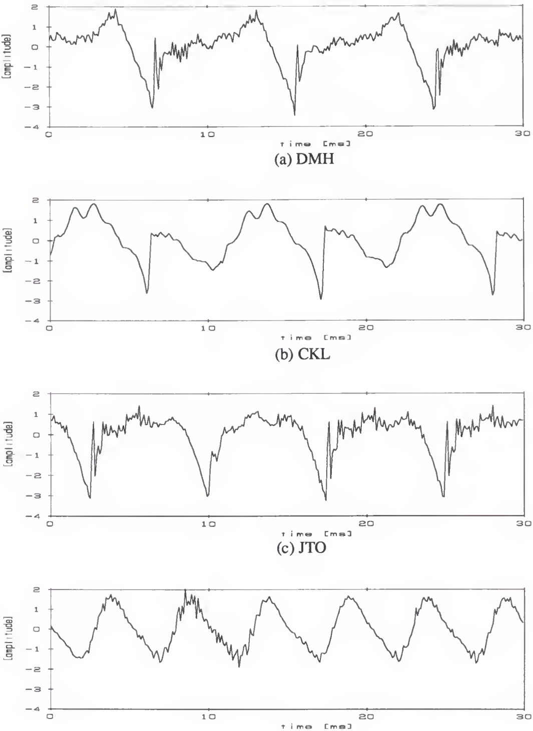
2.2.2 Glottal Source Model

As a result of the glottal inverse filtering algorithm, the glottal volume velocity or its differentiated waveform is obtained. Figure 2–2 shows several typical differentiated glottal flow waveforms obtained by glottal inverse filtering using the closed phase covariance method. The waveforms are quite different for the various types of voices (Childers and Ahn, 1995). The glottal pulse width is smaller for vocal fry voices. Breathy voices have large pulse widths, often making it appear that there is no closed phase. For all the voice types examined (i.e., modal, vocal fry, and breathy voices), the closing phase exhibits a steeper change of slope than the opening phase. Thus, the glottal flow waveforms are skewed to the right. Glottal pulse skewness also varies with voice type. It was also observed that for modal and vocal fry phonations the skewing was more apparent than for breathy phonations. Most of the modal and vocal fry voices show distinct closed phases. The closed phase is not always apparent for breathy voices, and in addition, multiple excitation can be found in vocal fry speech (Figure 2–2 (b)).

Several models were developed to describe this low frequency glottal volume velocity waveform (Rothenberg *et al.*, 1975; Rothenberg, 1981; Fant *et al.*, 1985; Milenkovic, 1993). Two of the most popular models are: 1) the polynomial model (Milenkovic, 1993), and 2) the LF model (Fant *et al.*, 1985). The glottal source modeling can be applied to various applications such as high quality speech synthesis, speech communication, and psychoacoustic studies of vocal quality.

2.2.2.1 Polynomial model

Polynomial fitting of the glottal volume velocity was recommended to allow for possible distortion in the data recording process (Milenkovic, 1993). Childers and Hu (1994) adopted a sixth order polynomial to model the differentiated glottal volume-velocity. The polynomial model can be written as



(a) DMH

(b) CKL

(c) JTO

(d) JMS

Figure 2–2. Glottal flow and normalized differentiated glottal flow waveforms for different type voices;
(a) : modal voice, (b) and (c) : vocal fry, (d) : breathy voice.

$$p(t) = C_0 + C_1\tau + C_2\tau^2 + C_3\tau^3 + C_4\tau^4 + C_5\tau^5 + C_6\tau^6 \ , \qquad 0 < t \leq T \tag{2–1}$$

where t is the time variable, $\tau = t \ / T$, and T is the pitch period. The coefficients, C_i, determine the shape of the differentiated glottal waveform, and are used as the parameters for the polynomial model. To derive the polynomial coefficients is an optimization problem, which is usually known as a polynomial fitting algorithm. The polynomial models can provide a robust and close match of the glottal waveform. Methods for determining the order of the polynomial is another important research issue.

2.2.2.2 LF model

Although the polynomial glottal source model is robust, the polynomial coefficients have little physical correlation to the movement of the glottis. To overcome this drawback, several acoustic models, whose parameters are closely related to the acoustic features of the glottal volume–velocity, have been suggested. These models are more applicable than the polynomial model for psychoacoustic studies. The LF model proposed by Fant *et al.* (1985) is one such model that has been widely used for glottal source modeling. This model describes the differentiated glottal flow rather than the glottal flow itself. The differentiated flow is commonly used in speech synthesis, and includes the effect of radiation at the lips. It is a good approximation for non–interactive flow parameterization in the sense that it ensures an overall fit to commonly encountered glottal flow wave shapes with a minimum number of parameters. Moreover, the existence of the residual closing phase in the LF model gives a flexibility for modeling a large variety of voice types efficiently.

The waveform of the LF model and its integral are shown in Figure 2–3. The LF model consists of two segments: exponentially growing sinusoid, and an exponential decaying function. Each segment is expressed as follows:

$$\frac{dU_g(t)}{dt} = E(t) = E_o \cdot e^{\alpha t} \sin \omega_g t \ , \qquad 0 \le t \le t_e \qquad (2\text{–}2)$$

$$E(t) = -\frac{E_e}{\varepsilon t_a}\left[e^{-\varepsilon(t-t_e)} - e^{-\varepsilon(t_c-t_e)}\right] , \qquad t_e \le t \le t_c \le T_0 \qquad (2\text{–}3)$$

where T_0 is the pitch–period interval within which a waveshape of the LF model is defined. At time t_e, both segments have the same value, E_e. Besides the above relationships, there is an area balance requirement, which keeps the DC level from drifting. Thus the integral of the LF model time–function through the entire pitch period should be zero, i.e.,

$$\int_0^{T_o} E(t) = 0 \ . \qquad (2\text{–}4)$$

The first segment of the LF model represents the differentiated flow from glottal opening to the instant when the main excitation occurs (the moment of maximum discontinuity in the glottal airflow function, which normally coincides with the moment of the maximum negative flow derivative). The three parameters of the first segment of the LF–model are:

(1) a scale factor E_o

(2) $\alpha = -B\pi$ where B is the “negative bandwidth” of the exponentially growing amplitude.

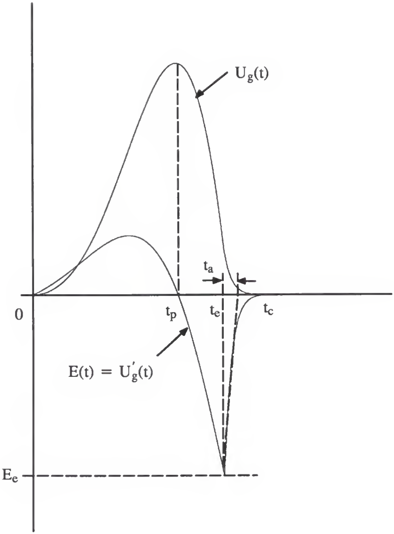


Figure 2–3. The LF–model of the differentiated glottal flow $U_g'(t)$ – not drawn to scale.

(3) $\omega_g = 2\pi F_g$ where $F_g = 1/2t_p$ and t_p is the rise–time (the time from glottal opening to maximum flow).

The second part of the model is an exponential segment that allows a residual flow (dynamic leakage), from the point of maximum closing discontinuity at time t_e towards maximum closure, when the vocal folds close at time t_c. The effect of the return phase on the source spectrum is, due to its exponential waveshape, approximately a first order low–pass filter with a cutoff frequency $F_a = 1/(2\pi t_a)$, where the parameter t_a is the time constant of the exponential curve and is determined by the projection on the time axis of the derivative at time t_e, at which the negative peak of the LF model occurs. (Fant and Lin, 1987). This means that the longer the return phase, the lower the cutoff frequency, and the greater the reduction of the high frequency energy. The parameter E_e is the negative amplitude of the excitation spike at time t_e. The parameter t_c is the moment when complete closure is reached. The parameter ε is the decay constant of the recovery phase exponential.

The LF model time–function is generated by using the direct synthesis parameters, i.e., E_o, α, ω_g, and ε. However, for many research applications, such as model fitting to inverse filtered glottal flow waveforms, it is easier to specify the timing parameters – t_p, t_e, t_a, t_c, and E_e rather than the direct synthesis parameters. The direct synthesis parameters can be easily computed from the timing parameters and E_e. The LF model timing/direct–synthesis parameters can be thought of as independent of one another, because a unique combination of timing/direct–synthesis parameters can generate a unique waveshape. The procedure to obtain the corresponding direct synthesis parameters from the timing parameters and E_e is as follows:

(1) The intermediate parameter ε can be determined by an iterative procedure from equation 2–3 by letting $t = t_e$, i.e., from

$$\varepsilon t_a = 1 - e^{-\varepsilon(t_c - t_e)} , \qquad (2\text{–}5)$$

For small values of t_a, ε is approximately equal to $1/t_a$.

(2) By definition, $\omega_g = \pi/t_p$.

(3) The solution for the parameter α can be obtained by applying the area balance constraint of equation 2–4 with the solution of E_o from equation 2–2

$$E_0 = -\frac{E_e}{e^{\alpha t_e} \, \sin \omega_g t_e} . \qquad (2\text{–}6)$$

In estimating the LF model parameters, the parameter t_c, which represents the closing instant, is usually set to T_o, the time of glottal opening for the following pulse period. This implies that the model may lack a closed phase. In practice, however, for small values of t_a the exponential function of the second part of the LF model will have negligible value, and, thus, provides an effective closed phase.

The LF–model function is continuous until the main excitation, and therefore does not introduce additional excitation at the flow peak. In comparison, Fant's model (Fant, 1979) consists of two different segments; a rising segment up to maximum flow and a falling segment down to complete closure. The discontinuity between the two segments introduces a secondary weak excitation at the flow peak. The major difference between these two models is that the LF model allows for a residual phase of progressive closure, while the Fant model generates an abrupt closure. The existence of the residual closing phase in the LF model gives the flexibility of modeling various voice types more efficiently.

Procedures have been proposed to fit the estimated glottal waveform from the glottal inverse filtering (GIF) with the LF model timing parameters (Gobl, 1988; Childers and Ahn, 1995). A major problem is identifying the glottal opening instant. Previous fitting procedures were designed either by interactively guessing the glottal opening instant or by using the EGG signal as auxiliary information (Childers and Krishnamurthy, 1985).

2.3 Analysis for Unvoiced Speech

Since voiced speech has a steady state segment of sufficient duration, determination of the vocal tract area function or articulatory vectors can be achieved using average values of the acoustic characteristics, i.e., the first four formant frequencies for the segment. The dynamics of the formants are not necessary for voiced speech and, therefore, most studies have been devoted to voiced speech. Compared to the inverse problem for voiced speech, much less has been done for the unvoiced speech. It is not known whether this is because the fricative spectra have insufficient information or because the analysis algorithms currently available for unvoiced speech are not adequate.

As one attempt at the inverse problem for unvoiced speech, Sorokin adopted an analysis–by–synthesis method to estimate the articulatory parameters, vocal tract shape and cross–sectional area function from fricative spectra. Minimum muscle work was used as a criterion in the optimization procedure. This criteria was used successfully in his prior work for determination of the vocal tract shape for voiced speech (Sorokin, 1992). However, for unvoiced fricatives, a proper initial approximation of articulator parameters was required to obtain an accurate and stable solution to the inverse problem (Sorokin, 1994).

In this research, a new analysis method for unvoiced speech is proposed. The proposed method is based on the human auditory model. We attempt to decompose the unvoiced speech spectrum into the vocal tract front cavity resonance and turbulence noise source spectrum. Estimation of the front cavity resonance characteristics are based on concepts from the human auditory model. In the next section, a review of concepts and terminologies for the auditory modeling will be given followed by a description of perceptual linear prediction (PLP) analysis algorithm which is based on the auditory model.

2.4 Human Auditory Model

The ideas from the human auditory model have been adopted in several speech analysis reports (Kewley–Port, 1982; Kewley–Port, 1983; Hermansky, 1986; Qi and Fox, 1992). This approach offers an alternative to other speech analysis methods. In this section, some basic concepts and ideas of auditory modeling are reviewed.

For several decades, research on the human perception of sound such as critical band rate, masking, loudness, and loudness level, has been a major research topic among psychoacoustic scientists. The final receiver of all sound sources is the human auditory system, which includes the outer ear, ear drum, ossicles, semicircular canals, cochlear, and auditory nerve. According to psychoacoustics, human perception of sound does not involve frequency, or energy, or power. "We do not perceive frequency. Rather we perceive pitch; we do not perceive level (amplitude), but loudness. We do not perceive spectral shape, modulation depth, or frequency of modulation; instead we perceive sharpness, fluctuation strength or roughness." (Zwicker and Zwicker, 1991)

2.4.1 Masking

One important aspect of hearing is the masking effect. The masking effect can be differentiated into simultaneous masking and nonsimultaneous masking. An example of simultaneous masking is when we have a conversation with friends while a train passes by. We need to stop our conversation for a while or increase the volume to continue our conversation. Another example can be found in an orchestral music. When loud instruments are played, the soft instruments are not audible because it is masked by the louder sounds. Nonsimultaneous masking is related to the duration of the masking noise.

2.4.2 Critical–band Rate Scale

The simultaneous masking effect can be understood if the *critical–band–rate* scale is used instead of the frequency scale. The concept of critical–band–rate scale is closely related to the physiology of the auditory system. It is directly related to the place along the basilar membrane where all the sensory cells are located in an equidistance configuration (Zwicker and Zwicker, 1991).

From previous experiments, below a frequency of about 500 Hz, the critical bandwidth is constant at about 100 Hz. Above that frequency, the critical bandwidth increases as a function of frequency. It is usually assumed that the relative bandwidth is 20% of the center frequency above 500 Hz. More exact values are given in Table 2–1 (Zwicker and Fastl, 1990). The audible frequency range up to 16 kHz can be divided into 24 critical bands. The scale is based on the fact that the human auditory system analyzes a broad spectrum into parts that correspond to critical bands A unit, *bark*, was defined for the critical–band rate. A bark is one critical–band width. If the bark scale is used, not only the masking effect but also many other effects such as pitch, just–noticeable frequency differences, and the growth of loudness as a function of bandwidth can be described more simply.

2.4.3 Loudness Level

The *loudness level* measure was created to characterize the loudness sensation of any sound. This measure is more precise than magnitude amplitude. The loudness level

Table 2–1. Critical band rate, lower(f_l) and upper limit(f_c) of critical bandwidth ΔF_G centered at F_c.

z (Bark)	f_l, f_u (Hz)	f_c (Hz)	z (Bark)	Δf_G (Hz)
0	0			
		50	0.5	100
1	100			
		150	1.5	100
2	200			
		250	2.5	100
3	300			
		350	3.5	100
4	400			
		450	4.5	110
5	510			
		570	5.5	120
6	630			
		700	6.5	140
7	770			
		840	7.5	150
8	920			
		1000	8.5	160
9	1080			
		1170	9.5	190
10	1270			
		1370	10.5	210
11	1480			
		1600	11.5	240
12	1720			
		1850	12.5	280
13	2000			
		2150	13.5	320
14	2320			
		2500	14.5	380
15	2700			
		2900	15.5	450
16	3150			
		3400	16.5	550
17	3700			
		4000	17.5	700
18	4400			
		4800	18.5	900
19	5300			
		5800	19.5	1100
20	6400			
		7000	20.5	1300
21	7700			
		8500	21.5	1800
22	9500			
		10500	22.5	2500
23	12000			
		13500	23.5	3500
24	15500			

is not only a sensational but also a physical value. It is defined as the sound pressure level (in dB) of a 1 kHz tone in a plane wave with frontal incident that is as loud as the sound. Its unit is the "phon." For example, if a 1 kHz tone has a 100 dB sound pressure level, than the loudness level is 100 phons. Using this measure, the loudness level for different frequencies of pure tones can be graphed. Lines that connect points of equal loudness in the hearing frequency range are called *equal–loudness contours*. Therefore, one can draw a number of equal–loudness curves for each sound pressure level. Figure 2–4 shows a typical equal–loudness curve. By definition, all curves must go through the sound pressure level at 1 kHz that has the same value in dB as the parameter of the curve in phons. For instance, the equal–loudness contour for 60 phons must go through 60 dB at 1 kHz. The dashed line indicates the threshold of quiet where the limit of loudness sensation is reached. In the frequency range above 200 Hz, shapes of the contour for all loudness levels are almost parallel to the threshold of quiet. At low frequency, however, the contour becomes shallow as the loudness level goes higher. The most sensitive range of threshold of quiet, between 2 kHz and 5 kHz, corresponds to a dip in the equal–loudness contours. For higher levels, the dip becomes even deeper, which means that tones in that frequency range are louder than expected from a parallel shift of threshold in quiet (Zwicker and Fastl, 1990).

2.4.4 Loudness

While the loudness level belongs somewhere between sensation value and physical value, *loudness* is a sensation value that corresponds most closely to the sound intensity of the test tone. This can be measured by comparing loudness (or softness) of a sound relative to a standard sound. Usually a 1 kHz tone with a level of 40 dB (1 sone) is used as the reference tone. Using the reference tone, the loudness function can be calculated and it often shows a power law as illustrated in Figure 2–5 by the solid line. The abscissa is the level of the 1 kHz tone and the ordinate is loudness on a logarithmic scale. The

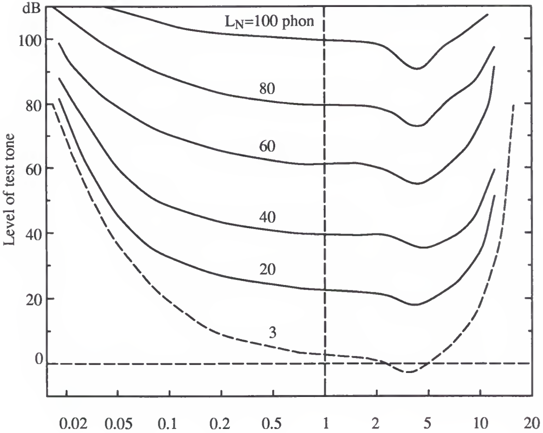


Figure 2–4. A typical equal–loudness contour for pure tones (after Zwicker and Fastl, 1990).

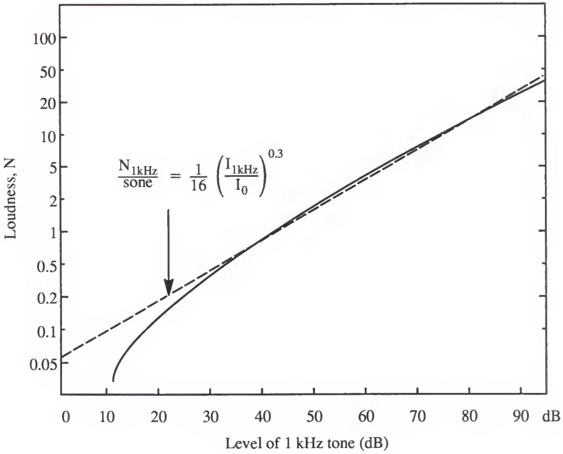


Figure 2–5. A typical intensity–loudness function. Loudness function of a 1 kHz tone (solid line) and its approximation (dashed line) (after Zwicker and Fastl, 1990).

curve can be approximated using power laws. The exponent of the power law that corresponds to a straight line is the steepness of the straight line for levels above 30 dB. Usually, the exponent is 3/10 = 0.3 for the sound pressure level of 30 dB or above. At levels below 30 dB, the approximation is no longer valid.

2.5 Perceptual Linear Prediction (PLP)

2.5.1 Introduction

The concept of the auditory model has been widely applied to speech coding and data compression. In speech analysis there are several analysis methods based on the auditory model of human hearing (Kewley–Port and Luce, 1984; Kurowski and Blumstein, 1984, 1987; Qi and Fox, 1992). Perceptual linear predictive (PLP) analysis is one of these algorithms and it is introduced in this section.

The basic idea of the PLP analysis technique, originally suggested by Hynek Hermansky (Hermansky, 1990), is to approximate the auditory spectrum of speech by an all–pole model. The auditory spectrum is obtained from the speech signal by filtering using a critical–band filter bank followed by an equal loudness curve pre–emphasis and an intensity to loudness conversion by the intensity–loudness power law. Then, the auditory spectrum is modelled by an autoregressive (all pole) model.

The PLP analysis yields an auditory spectrum with nonuniform frequency resolution. It has low resolution in the low frequency range and high resolution in the higher frequency range. It has been shown in speech recognition experiments that the low order PLP (5th order) has a performance equivalent to a 13th order conventional LP model (Hermansky, 1990).

2.5.2 Algorithm

The block diagram of the PLP algorithm is shown in Figure 2–6. Steps 1 to 3 are to obtain an estimate of the auditory spectrum based on the human auditory model. The steps 4 and 5 are for the all pole modeling of the auditory spectrum. The optional steps 6, 7, and 8 are to transform the model spectrum back to the original amplitude–frequency domain. This algorithm is implemented in the Matlab® language.

2.5.2.1 Spectral analysis

First, the speech to be analyzed is segmented and weighted by the Hamming window

$$W(n) = 0.54 + 0.46 \cos \left[\frac{2\pi n}{(N-1)} \right], \tag{2–7}$$

where N is the length of the window.

The fast Fourier transform (FFT) is used to transform the windowed speech segment into the frequency domain. The real and imaginary components of the short–term speech spectrum S(ω) are squared and added to get the short–term power spectrum

$$P(\omega) = \Re [S(\omega)]^2 + \Im [S(\omega)]^2. \tag{2–8}$$

Then, the power spectrum P(ω) is warped along its frequency axis ω onto the Bark scale using the approximation given by

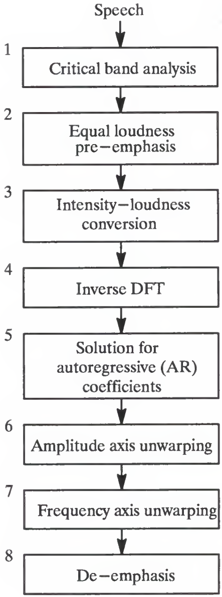


Figure 2–6. Block diagram of the PLP speech analysis method.

$$\Omega(\omega) = 6 \cdot \log [\frac{\omega}{1200\pi} + \sqrt{ (\frac{\omega}{1200\pi})^2 + 1 }], \qquad (2\text{–}9)$$

where ω is the angular frequency in rad/s (Schroeder, 1977).

2.5.2.2 Critical band analysis

Next, the critical band filter is simulated by summing the weighted short–time spectrum. The warped power spectrum of the speech P(Ω) is then convolved with this simulated critical band masking curve Ψ(Ω) yielding samples of the critical band power spectrum

$$\Theta(\Omega_i) = \sum_{\Omega=-1.3}^{2.5} P(\Omega - \Omega_i) \cdot \Psi(\Omega). \qquad (2\text{–}10)$$

where the critical–band curve is given by

$$\begin{aligned} \Psi(\Omega) &= 0 && \text{for } \Omega < -1.3, \qquad (2\text{–}11)\\ &= 10^{\,2.5\cdot(\Omega + 0.5)} && \text{for } -1.3 \le \Omega < -0.5,\\ &= 1 && \text{for } -0.5 \le \Omega < +0.5,\\ &= 10^{\,-1.0\cdot(\Omega - 0.5)} && \text{for } 0.5 \le \Omega < +2.5,\\ &= 0 && \text{for } \Omega \ge 2.5\ . \end{aligned}$$

2.5.2.3 Equal loudness pre–emphasis

The sampled Θ(Ω(ω)) is preemphasized by the simulated equal–loudness curve in order to compensate for the non–equal perception of loudness at different frequencies.

$$\Xi(\Omega(\omega)) = E(\omega) \cdot \Theta(\Omega(\omega)), \tag{2–12}$$

where the equal–loudness curve is given by

$$E(\omega) = \frac{(\omega^2 + 56.8 \times 10^6)\,\omega^4}{(\omega^2 + 6.3 \times 10^6)(\omega^2 + 0.38 \times 10^9)(\omega^6 + 9.58 \times 10^{26})}, \tag{2–13}$$

which is an approximation to the nonequal sensitivity of human hearing at different frequencies (Hermansky, 1990).

2.5.2.4 Intensity–loudness conversion

According to the power law of hearing (Stevens, 1957; Zwicker and Fastl, 1990), a cubic–root amplitude compression is performed by the following relation

$$\Phi(\Omega) = \Xi(\Omega)^{\frac{1}{3}}. \tag{2–14}$$

This is to simulate the nonlinear relation between the intensity of speech sound and its perceived loudness. This operation also reduces the spectral–amplitude variation of the critical band spectrum so that the following all–pole modeling can be done by a relatively low model order. Although this approximation is not true for very loud or very quiet sounds, it is a reasonable approximation for speech sounds.

2.5.2.5 AR modeling

Finally, $\Phi(\Omega)$ is modelled by an all–pole model using the autocorrelation method. The inverse DFT is applied to $\Phi(\Omega)$ to get autocorrelation values and the coefficients of the all–pole model can be obtained by solving the Yule–Walker equations using the autocorrelation values.

2.5.3 Characteristics of the PLP Algorithm

It is well known that the linear prediction spectrum analysis approximates the spectral envelop equally well at all frequencies of the analysis band. However, the human auditory system is inconsistent with the property of linear prediction analysis. In fact, the spectral resolution of human hearing decreases with frequency beyond about 800 Hz. For typical amplitude levels, hearing is more sensitive in the middle frequency range of the audible spectrum.

In comparison with conventional linear prediction analysis, the PLP spectrum analysis is more consistent with human hearing. The PLP analysis yields an auditory spectrum with relatively low frequency resolution. Furthermore, the frequency resolution of the auditory spectrum is nonuniform. In the higher frequency range, it has less resolution, which is in agreement with the characteristics of the human auditory system. The effective second formant F_2' and the 3.5–Bark spectral–peak integration theories of vowel perception are well accounted for in the PLP analysis. The algorithm is computationally efficient and yields a low–dimensional representation of speech.

2.6 Effective Second Formant F_2' and Front Cavity Resonance Frequencies

In the study of auditory perception, it is observed that two spectral peaks are all that are needed for simulating front vowels and that one spectral peak is sufficient for simulating back vowels (Delattre *et al.*, 1952). Carlson *et al.* (1975) and Bladon and Fant (1978) also found that vowels can be simulated perceptually by using only two spectral peaks. While the first peak is held at the first formant F1, the second spectral peak has to be put between F_3 and F_4 for simulating high front vowels, such as /i/, and into the vicinity of the second formant F_2 when simulating back vowels, such as /a/. Fant called the second spectral peak the effective second formant F_2' because it does not correspond to any of the formants (Fant and Risberg, 1962).

It seems that the F_2' has some correlation with speech production as well as speech perception. Fant showed that, when simulating a vowel by a single harmonic signal, the listeners response peaked close to the uncoupled front cavity of the vocal tract. He suggested that the F_2' might be equivalent to this resonance frequency of the vocal uncoupled front cavity. According to the basic acoustic theory of speech production, the fundamental resonance frequency of the front cavity may be associated with any of the first four formants (Fant, 1960).

An experiment by G. M. Kuhn (1975) using spectrographic data from two types of speech, one from normal speech and the other from fricative speech, suggested that it may be possible to estimate the frequency of the front cavity resonance from information in the speech signal. It is also suggested that the front cavity resonance could be estimated from speech data that is of a form more like that found in the auditory system (Kuhn, 1975).

Formulas for calculating F_2' as a function of the first four formant frequencies have been proposed. Carlson *et al.* (1975) designed a formula to predict a continuous shift of F_2' between the extreme frequencies of F_2 and $\sqrt{F_3F_4}$, taking into account relations between formant frequencies and spectrum shape. The formula is

$$F_2' = \frac{F_2 + c(F_3F_4)^{1/2}}{1 + c} , \qquad (2\text{–}15)$$

where

$$c = \left(\frac{F_1}{500}\right)^2 \left(\frac{F_2 - F_1}{F_4 - F_3}\right)^4 \left(\frac{F_3 - F_2}{F_3 - F_1}\right)^2 . \tag{2–16}$$

It is reported that the values of F2′ predicted by this formula are within 75 Hz, on the average, of values predicted by a model of the cochlea for nine Swedish vowels (Carlson *et al.*, 1975). An improved version of the formula has been developed and tested with a set of cardinal vowels (Bladon and Fant, 1978). The new formula is based on a spectrum prominence model, which postulates interdependencies between formant frequencies and spectrum levels (Fant, 1960). The new formula is

$$F_2' = \frac{F_2 + c^2(F_3F_4)^{1/2}}{1 + c^2} , \tag{2–17}$$

where

$$c = K(f) \cdot \frac{A_{34}}{A_2} . \tag{2–18}$$

The ratio between A_{34} and A_2 can be approximated as

$$\frac{A_{34}}{A_2} = \frac{B_2 \cdot F_2 \, (1 - F_1^2/F_2^2) \, (1 - F_2^2/F_3^2) \;\; (1 - F_2^2/F_4^2)}{(F_4 - F_3)^2 \, (\frac{F_3F_4}{F_2^2} - 1)} , \tag{2–19}$$

where K(f) = 12 and B2 = 100 Hz as initial approximations. The calculated effective second formant F_2' of most vowels followed the second formant F_2 within an average of 7 Hz. However, both of the above two formulas fail to produce a reliable estimate of the F_2' for all of the cardinal vowels.

A different approach to estimate F2′ is the auditory based analysis method, which provides a new solution to the problem. The perceptual linear prediction (PLP) technique described in the previous section shows that the second spectral peak in the PLP spectrum is well matched with the effective second formant. This suggests that the PLP algorithm can be used to estimate the resonance frequency of the front cavity since the effective formant provides the data to calculate resonance frequency and bandwidth for the front cavity (Hermansky and Broad, 1989). In the next chapter, a research design based on the estimation of the front cavity resonance using the PLP technique is introduced.

CHAPTER 3
RESEARCH DESIGN

Speech sounds can be classified into 3 categories based on the mode of excitation: voiced sounds, unvoiced sounds and plosives. Since most portions of speech are voiced sounds, analyzing voiced sounds provides us with an understanding of the production of phonetic information and vocal characteristics. For example, using conventional glottal inverse filtering algorithms, the air flow at the glottis can be estimated as well as the vocal tract transfer function. The waveshape of the glottal volume velocity is a major factor that determines the vocal quality of the speech (Childers and Lee, 1991; Childers and Ahn, 1995). The vocal tract transfer function can be estimated by the linear prediction algorithm or other spectral estimation algorithms (Kay, 1988), which determines the phonetic information of the speech.

On the other hand, the acoustics and aerodynamics involved in the production of unvoiced speech are far from being completely understood. Estimating the vocal tract transfer function and the waveshape at the glottis is not enough to determine the unvoiced noise source characteristics because unvoiced speech is usually generated at a constriction within the vocal tract. To be more precise, the noise source may be located in the vocal tract for fricatives, while the glottis is the noise source for aspiration noise /h/. There are many unvoiced speech sounds that are generated by both types of constrictions: one at the glottis and another in the vocal tract (Stevens, 1971). One purpose of this research is to understand the production of unvoiced speech and propose an unvoiced speech generation model.

3.1 Research Overview

Unvoiced speech can be divided into 3 groups according to the speech generation mechanism, i.e., unvoiced fricatives, unvoiced stops, and affricates. Aspiration noise can be regarded as unvoiced speech as well.

An unvoiced fricative is generated by exciting the vocal tract by a steady air flow, which becomes turbulent near the constriction where the velocity of the airflow increases due to the reduced cross–sectional area of the constriction. The location of the constriction determines which sound is produced. The constriction separates the vocal tract into two cavities: front cavity and back cavity. The unvoiced speech sound is generated from the front cavity and the back cavity traps energy and, thereby, introduces anti–resonances into the vocal output.

An unvoiced stop consonant is produced by forming a complete closure in the vocal tract and then releasing that closure. The closure is formed by a particular articulator, i.e., the lips, the tongue blade or the tongue body, and then released by moving the articulator rapidly. The initial rapid increase of the cross–sectional area at the constriction gives rise to a transient, and, there is a brief burst of turbulence noise following the transient.

An unvoiced affricate is a dynamical sound that can be modelled as a concatenation of the unvoiced stop and the fricative consonant. Therefore, this sound has the characteristics of both stop and fricatives consonants.

An aspiration is produced by a turbulent air flow at the glottis with little or no vibration of the vocal folds. The characteristics of the aspiration /h/ are similar and dependent to the vowels that follow the aspiration noise because the vocal tract shape is in the position for the following vowel during the production of the aspiration noise. Since the noise source, constriction, is located at the glottis, there is no front and back cavity in aspiration.

The four unvoiced sound groups are fricative, affricate, stop, and aspiration. The fricative and aspiration are generated by a relatively steady flow of air. Therefore, it is much easier to analyze fricatives and aspirations than other unvoiced consonant categories. The goal of this research is to investigate the properties of the vocal tract transfer function in relation to source location and the vocal tract cavity structure. In order to accomplish this goal, we need to be able to estimate the resonance frequencies of the front cavity and the back cavity of the vocal tract transfer function. The effect of the back cavity resonance on the overall spectrum may be negligible if the constriction is sufficiently narrow and long. (Heinz and Stevens, 1961).

A research procedure based on the analysis–by–synthesis method is designed as shown in Figure 3–1 to study the production mechanism of the unvoiced consonants. In the analysis phase, the turbulence noise spectrum and the vocal tract front cavity transfer function is estimated. The front cavity resonance frequency can be estimated by PLP analysis and the turbulence noise source spectrum can be obtained by inverse filtering. Since the front cavity characteristics from the constriction to the lips are accounted for by the front cavity transfer function, removing the front cavity resonance from the fricative spectrum results in the turbulence noise source spectrum. Once the noise source spectrum is obtained by inverse filtering, it can be used to establish the relationship between the noise source spectrum and the constriction in the vocal tract. Also, the front cavity resonance can be used to estimate the front cavity length. The narrow tube model parameter can be directly mapped to the articulatory parameters from the glottis to the constriction. One can use the speech inverse filtering procedure to obtain the articulatory parameters for the vocal tract front cavity (Hsieh, 1994). Estimating the articulatory parameters, i.e., the position vectors of the articulators, is not in the scope of this research.

Based on the estimated parameters, synthetic speech based on an articulatory synthesizer developed by Hsieh (1994) is generated, and is evaluated by informal listening tests. The listening tests are mainly focused on the intelligibility of the synthetic

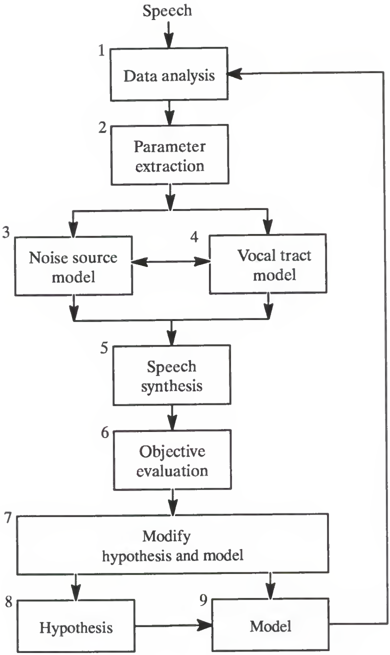


Figure 3–1. Research design.

unvoiced speech. The relations between analysis parameters and the type of the synthetic speech will be studied. The basic idea is that the center frequency of the turbulence noise source spectrum is closely related to the location of the source, i.e., the location of the constriction in the vocal tract. The center frequency in the source spectrum can be used to calculate the location of the source in the vocal tract and, furthermore, can be applied to classify the unvoiced speech. The analysis results for the center frequency closely matches previously reported simulation results, thereby, validating the analysis procedure.

To understand factors that control the unvoiced speech generation process, we shall describe an unvoiced fricative production model, which includes the turbulence noise generator and the vocal tract structure. Details of the speech analysis and the synthesis methods that are used in this research will also be described later in this chapter.

3.2 Unvoiced Fricatives Production

Unfortunately, an aerodynamic study of turbulence noise is not well established and often based on empirical rules. When air flows through a constriction, velocity of the flow increases due to the small cross–sectional area of the constriction. A speech sound generated in this manner is called turbulence noise. In this mode of speech generation, the air flow in the vocal tract is determined by two constrictions: one at the glottis and another one above the glottis formed by the tongue or lips. When the cross–sectional area of the supraglottal constriction is small compared to the glottal opening, the supraglottal constriction plays a major roll in generating the turbulence. A turbulence noise generated in this vocal tract configuration is called frication noise. On the other hand, if the supraglottal constriction area is large, then the air flow is independent of the supraglottal constriction and noise generated in this configuration is called aspiration noise.

3.2.1 Model for Turbulence Noise Generation for Fricative Consonants

A schematized model of the vocal tract for an unvoiced fricative consonant is shown in Figure 3–2. A_g and A_c are the cross–sectional areas of the glottis and the constriction, respectively. Likewise, L_g and L_c are length of the glottis and the vocal tract constriction, respectively. If $A_g > A_c$, the supraglottal constriction plays a major roll in the generation of fricative speech. Aspiration noise is generated when $A_g < A_c$.

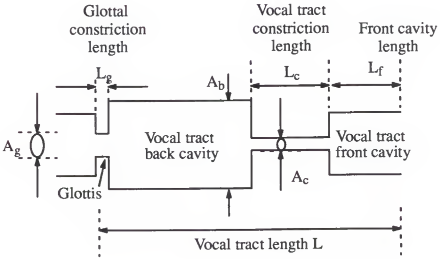


Figure 3–2. A model of the vocal tract for a fricative consonant. The vocal tract constrict divides the vocal tract into front and back cavity. A_c and L_c determine the characteristic of turbulence noise source.

As explained in Figure 3–2, when the noise source is in the vocal tract, the constriction separates the vocal tract into two cavities: front cavity and back cavity. Speech is radiated from the front cavity, while the back cavity serves to trap energy, and, thereby, introduces anti–resonances into the vocal output. It is reported that the back cavity resonances may have negligible effect on the fricative spectrum if the constriction is sufficiently long and narrow (Heinz and Stevens, 1961). Therefore, the most prominent features of the fricative spectrum are determined by the turbulence noise source at the constriction and by the resonances of the oral cavity in front of the constriction, i.e., the front cavity resonance frequencies.

Unvoiced fricatives are produced by a turbulence noise source near a constriction either at the glottis or above the glottis depending on the cross–sectional areas, A_g and A_c. The location of the constriction, i.e., the place of the articulation, determines the sound produced. For fricative noise, in which the constriction is formed by the tongue body and/or tongue tip, the length of the constriction, L_c, is typically a few centimeters long. For aspiration noise, the length of the constriction formed by the glottis, L_g, may be only a few millimeters, which is the average width of the glottis. The constriction in the vocal tract may be abrupt or gradual depending on the articulatory position. It has been found experimentally that the pressure drop across the constriction is proportional to the square of the particle velocity, $V_c{}^2$, and the pressure drop across the constriction Δp was approximated by the Bernouilli equation

$$\Delta p = k\varrho\frac{V_c^2}{2} = k\varrho\frac{U^2}{2A^2} , \qquad (3\text{–}1)$$

where k is a constant, ϱ is the density of the air, V_c is the flow velocity in the constriction, U is the volume velocity in the constriction and A is the cross–sectional area of the constriction. The constant k is dependent on the ratio of the cross–sectional area A_b and A_c. It is also dependent on the rate of contraction and expansion of the constriction. Stevens found a value of 0.9 a reasonable average for constrictions that can be found in the vocal tract of normal speakers (Stevens, 1971).

The configuration of the vocal tract for the generation of an unvoiced turbulence noise source can be modelled as in Figure 3–3. This model was originally suggested by Fant (1960) who used it in calculations for various fricatives and stop consonants. The source was assumed to be independent of the constriction area A_c. The impedance, Z_2, is

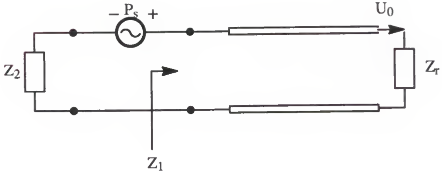


Figure 3–3. Equivalent circuits of the turbulence noise generation.

seen looking upstream from the sound pressure source, P_s, and is frequency dependent. The impedance looking downstream from the pressure source is Z_1. According to this model, the front and back cavity resonance frequencies are determined by Z_1 and Z_2, respectively. The volume velocity U_0 at the mouth flows through the radiation impedance Z_r. The transfer function U_0/P_s, i.e., the transfer function from the pressure source to the volume velocity at the lips can be calculated as

$$\frac{U_0}{P_s} = \frac{1}{Z_1 + Z_2}, \qquad (3\text{–}2)$$

Therefore, due to the frequency dependent impedance Z_2, the spectrum of the turbulence noise is characterized by poles at the natural frequencies for which $Z_1 + Z_2 = 0$, and by zeros at the frequencies for which $Z_2 = \infty$. As discussed above, the most important two components in this model are the turbulence noise source, P_s, at the constriction and the impedance Z_1 looking downstream from the pressure source that determines the front cavity resonance frequencies.

3.2.2 Noise Source Model

The series pressure source P_s in Figure 3–3 can be modelled using a spoiler in a semi–infinite tube as in Figure 3–4. The airflow in a narrow tube generates turbulence noise and the intensity and spectrum of the noise are decided by the Reynolds number Re. If $R_e > R_e^*$, where R_e^* is the critical Reynolds number, then a noise with high intensity is generated, which is a turbulence noise. The Reynolds number is defined as

$$R_e = \frac{\varrho_0 v h}{\mu}, \qquad (3\text{–}3)$$

where ϱ_0 is air density, v is linear velocity of air flow, h is characteristic dimension of the constriction and μ is viscosity of air (Sorokin, 1985).

It has been found experimentally that the spectrum of the series pressure source is relatively flat over a frequency range of two or three octaves centered on 0.2 V/D, where V is the velocity and D is a characteristic dimension (Stevens, 1971). More specifically, the center frequency can be represented as

$$f_c = 0.2 \cdot \frac{V}{A^{3/2}}, \qquad (3\text{–}4)$$

where, V is the volume velocity of the air flow near the constriction and A is the cross sectional area of the constriction. For typical values of the volume velocities and constriction sizes encountered in turbulence noise generated speech, the center frequency is in the range of 500 – 3000 Hz. The lower end of this range is for aspiration noise /h/ and the higher end is for fricative noise. In other words, as the position of the constriction

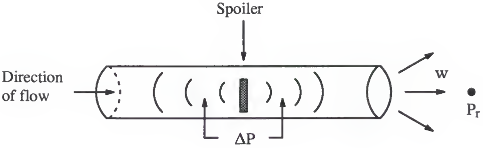


Figure 3–4. A spoiler in an otherwise uniform semi–infinite tube with airflow. Pressure drop across the spoiler is ΔP; and P_r is the sound pressure at a distance from the opening (Stevens, 1971).

moves forward from the glottis to the lips, the pole of the transfer function moves higher in frequency, e.g., from about 500 Hz for aspiration noise /h/ to about 4000 Hz for fricative noise [s] (Stevens, 1971).

Under the plane wave assumption, the sound pressure of turbulence flow can be taken as proportional to the square of the volume velocity of the airflow and inversely proportional to the constriction area, A_c (Stevens, 1971). The location of the turbulence noise source may be located at the center of, or immediately downstream from the constriction region, or possibly at a combination of these places, or spatially distributed along the constriction region (Fant, 1960; Stevens, 1971, 1993a, 1993b; Lin, 1990).

Basically, the noise source model defines the characteristics of the noise source as a function of the airflow through the constriction and of the constriction cross-sectional area, A_c. Meyer-Eppler (Broad, 1977b) found that the rms sound pressure, P_{rms}, of the noise could be expressed as

$$P_{rms} = A_c(R_e^2 - R_e^{*2}) \;, \qquad (3\text{–}5)$$

where R_e is the Reynolds number and R_e^* is the critical Reynolds number. Fant (1960) adopted a serial noise pressure source and reformulated the P_{rms} as a function of the pressure drop through the constriction and the effective width of the constriction. Lin (1990) extended Fant's model to include the frictional and turbulent losses inside the constriction. Both Fant and Lin tried to reconstruct the fricative spectra from area functions by using the acoustic transfer function. However, some fricatives have been modeled quite successfully and some are unsatisfactory (Badin, 1989, 1991).

Klatt (1980) used a random number generator, a spectrum-shaping filter, and an amplitude modulator to model the turbulent flow for the formant synthesizer. The spectrum-shaping filter was to simulate the spectral characteristics of the turbulent flow. A first order IIR filter was used to obtain the volume velocity due to a random pressure source. Childers and Lee (1991) have used a FIR filter to model highpass-filtered turbulence noise.

By including a latent random pressure source, P_n, and an inherent constriction loss, R_n, in each elemental section of the vocal tract, Flanagan and Cherry (1968) could introduce automatically the turbulent flow excitation at any section. However, as Sondhi and Schroeter (1987) pointed out, the Flanagan and Cherry (1968) model did not produce satisfactory unvoiced sounds due to the too high "back" cavity impedance. Sondhi and Schroeter (1986, 1987), thus, modified the model into a parallel flow source $U_n = P_n/R_n$, which was located downstream from the constriction. The P_n is given by

$$\begin{aligned} P_n &= \text{turbg} \cdot \text{rand} \cdot \left(R_e^2 - R_{ec}^2\right), && \text{for } R_e > R_{ec} \\ &= 0, && \text{for } R_e \leq R_{ec} \qquad (2.27) \end{aligned}$$

where turbg is empirically determined as the turbulence gain, and rand is a random number uniformly distributed between –0.5 and 0.5. A first-order IIR filter with cutoff frequency 2000 Hz was used to lowpass the flow. Figures 2–21(a) and (b) show the equivalent circuits of the serial and parallel turbulence sources, respectively.

We adopt the turbulence noise source model from Sondhi and Schroeter (1986, 1987). However, our model allows the user to place the turbulence noise source at the center of, or immediately downstream or upstream from the constriction region, or spatially distributed along the constriction region. The turbulence gain and critical Reynolds number can also be specified.

According to the above discussion, the speech generation model for unvoiced fricatives, affricates can be summarized as in Figure 3–5.

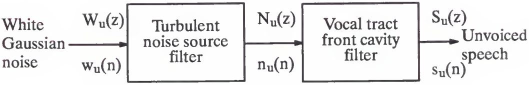


Figure 3–5. A model of unvoiced speech generation.

3.2.3 Quarter–wave Resonance

In the previous chapter, we explained that the resonance of the vocal tract front cavity closely represents the effective second formant. From experiments using fricative speech by Kuhn (1975), there is a component in the spectrum of fricative speech, which can be interpreted to be the quarter–wave resonance of the front cavity. The component was consistently close to the effective second formant, i.e., the front cavity resonant frequency. If the front cavity resonance frequency can be estimated from the speech signal, the front cavity length can be calculated using the formula

$$l = \frac{c}{4f} , \qquad (3\text{–}6)$$

where c is the speed of sound (353m/sec for 35°) and f is a quarter–wave resonance. For example, a quarter–wave resonance at 700 Hz corresponds to the front cavity length of 12.6 cm. At 3000 Hz, the length would be about 2.9 cm. Consequently, if one can estimate the effective second formant, which is believed to be a good estimate of the vocal tract front cavity resonance frequency, the length of the front cavity can be estimated directly from speech using the above equation. Estimation of the front cavity resonance frequency will be explained in the next section.

3.2.4 Estimation of the Vocal Tract Front Cavity Resonance Frequency

In this research, the PLP algorithm was used to estimate the effective second formant F_2', which might be equivalent to the resonant frequency of the vocal tract uncoupled front cavity (Fant, 1978; Bladon and Fant, 1978; Hermansky and Broad, 1989; Hermansky, 1990). In this section, examples of tracking the effective second formant using the PLP algorithm will be illustrated. The examples show that the F_2' estimated using the PLP is close to the second formant for back vowels and close to the third formant for front vowels.

The PLP algorithm is implemented in Matlab® Language that can be run either in the PC environment or in the Unix® operating system environment without changing the source code. The PLP algorithm is applied to sustained vowels and the first and second spectral peaks were estimated from the PLP spectrum. The estimates of the F_2' from cardinal sustained vowels are compared with the perceptually related formants determined by Bladon and Fant (1978) in Table 3–1. Note the numerical values of the table is in the scale of Bark, and the Bark scale can be transformed back to a linear scale using the relationship:

$$f_f = 600 \cdot \sinh(\frac{f_b}{6}) \ , \qquad (3\text{–}7)$$

where f_f is the frequency in Hz and f_b is the frequency in Bark (Schroeder, 1977). It can be observed that the estimates using the PLP analysis are close to the known values for the vowels.

As another example of applying the PLP algorithm to continuous speech, a sentence “We were away a year ago” spoken by a male speaker with a normal vocal quality is analyzed using the PLP algorithm. For reference, the spectrogram of the input sentence is shown in Figure 3–6. It is a wide band spectrogram with a frame length of 12.8 msec windowed using the Hamming window. The number of overlap points was 56 points. One can observe that the first formant is slightly above 400 Hz. The second formant ranges from 500 Hz to 2000 Hz while the third formant ranges from 1500 to 3000 Hz.

First, a formant track of the whole sentence is estimated in order to compare with the result of PLP analysis. Figure 3–7 is the formant track using an iterative glottal inverse filtering algorithm called a pitch synchronous iterative adaptive inverse filtering (PSIAIF) based on an iterative adaptive inverse filtering (IAIF) algorithm

Table 3–1. Comparison of the perceptually estimated formants based on the two formant model (Bladon and Fant, 1978) and PLP estimates. The values are given for the Bark scale.

	Perceptual		PLP		Error	
	F_1	F_2	F_1'	F_2'	$F_1'–F_1$	$F_2'–F_2$
/a/	5.7	8.2	5.9	Merged	0.2	N/A
/i/	2.9	14.1	3.7	12.6	0.8	–1.5
/u/	2.8	5.8	4.3	Merged	1.5	N/A
/e/	4.3	12.5	4.7	11.2	0.4	–1.3
/ɔ/	5.1	6.6	5.8	Merged	0.7	N/A
/ʌ/	5.4	9.0	5.5	Merged	0.1	N/A

(Alku, 1992). In the PSIAIF method, the glottal pulse is computed by applying the IAIF to the same speech signal twice. The first analysis gives a result for a glottal excitation that spans several pitch periods, which will be used to determine positions and lengths of frames for pitch synchronous analysis. The final result for the glottal waveform is obtained by analyzing the original speech pitch by pitch. One can find that the formant track is well matched with the spectrogram except for several false detections. These false detections can removed by further processing this formant track using a smoothing algorithm or applying rules from phonetics. Figure 3–8 is an estimate of the effective second formant tract using the PLP algorithm overlapped with the formant tract of Figure 3–7. Estimates of the first and second peaks in the PLP spectrum are drawn with circles. It can be observed that the F_2' follows the second formant for the back vowels such as /o/ and /u/. For the front vowels, such as /i/, the F_2' is somewhere between the second formant, F2, and the third formant, F3.

In summary, the effective second formant does not appear to be linked to any particular formant (Hermansky and Broad, 1989). Instead, it can be regarded as a combination of formants. The effective second formant is close to the vocal tract front cavity resonance and it can be estimated using the PLP algorithm. The front cavity resonance can be utilized to estimate the length of the vocal tract front cavity. Also, the spectrum of the turbulence noise source can be estimated by inverse filtering the vocal tract front cavity resonance from the speech signal.

In the next sections, a manual inverse filtering tool for manipulating the transfer function of the vocal tract by controlling the first five formant frequencies and their bandwidths will be explained. Some examples of manual inverse filtering, as well as selective inverse filtering will be described.

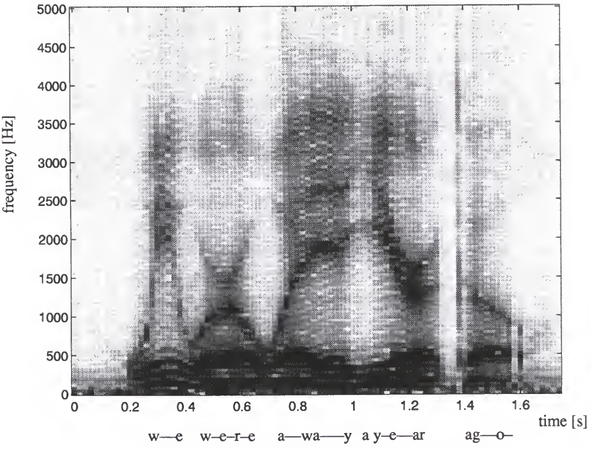


Figure 3–6. A spectrogram of the sentence "We were away a year ago."

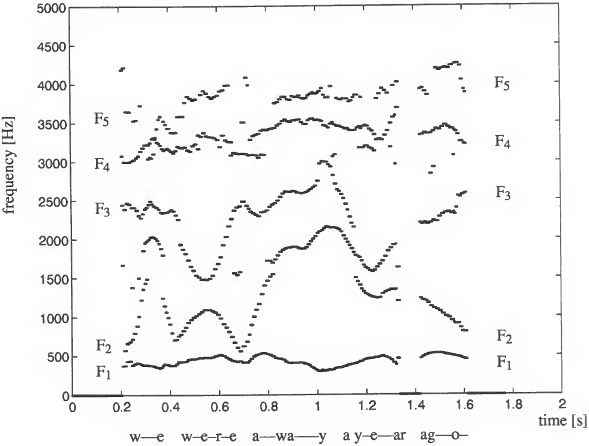


Figure 3–7. Formant tracks for the sentence “We were away a year ago” using the analysis tool.

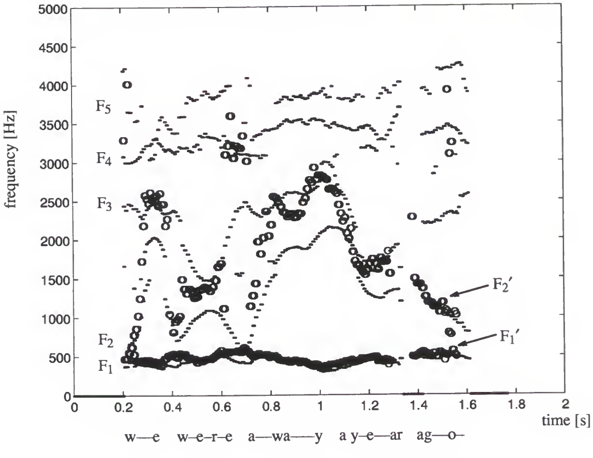


Figure 3–8. Formant tracks (drawn in dots) for the sentence "We were away a year ago" overlapped with the effective formants F_1' and F_2' (drawn in circles) estimated using the PLP algorithm.

3.3 Manual Inverse Filtering and Selective Inverse Filtering

As reviewed in section 2.2.1, many automatic glottal inverse filtering algorithms help speech researchers to understand the characteristics of the glottal volume velocity. (Childers and Lee, 1991). However, current automatic algorithms work well for only limited data, i.e., most algorithms do not give reliable results except for a normal male voice. More robust and reliable automatic algorithms that work for a female voice and pathological voices must be developed. In order to circumvent this problem and assist the analysis of a more wide variety of speech types, a manual inverse filtering algorithm is implemented here.

3.3.1 Manual Inverse Filtering

The vocal tract model consists of 5 digital resonators connected in cascade. Each resonator is for a formant in the vocal tract transfer function and can be represented as a second–order difference equation, with a z–transform given as

$$T(z) = \frac{A}{1. - Bz^{-1} - Cz^{-2}} , \qquad (3\text{–}8)$$

where $z = e^{j2\pi fT}$, j is an imaginary number $\sqrt{(-1)}$, and f is frequency in Hz. The coefficients A, B, and C are

$$\begin{aligned} C &= -e^{-2\pi f_b T} , && (3\text{–}9) \\ B &= -2e^{-\pi f_b T} \cos(2\pi f_f T) , \\ A &= 1 - B - C , \end{aligned}$$

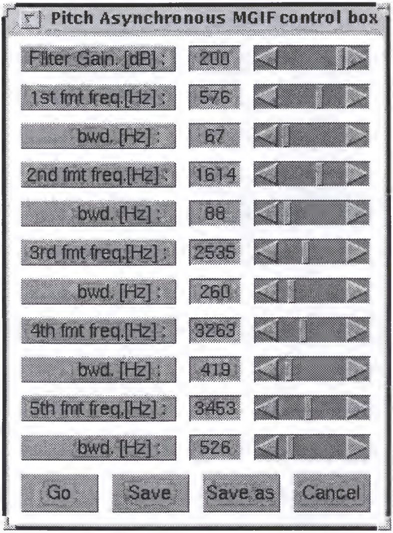
where f_b is a desired formant bandwidth, f_f is a desired formant frequency, and T is a sampling frequency. The vocal tract transfer function consists of 5 resonators that are controlled by 5 formant frequencies and bandwidth pairs. The transfer function of the inverse filter is the reciprocal of the vocal tract transfer function, in which the formant frequency and bandwidth can be adjusted manually. Figure 3–9 shows the control window setting for numerical values of the formant frequency and bandwidth. By moving the slide bar or typing the desired numerical value in the box, the formant frequency and bandwidth of the vocal tract model can be set. Then, the preemphasized input speech is filtered by the inverse filter and FFT spectrum of the filter's output, which is the differentiated glottal volume velocity, is displayed in a separate window. The coefficients of the inverse filter can be adjusted in order that the spectrum of the cascade vocal tract model is well matched to the spectrum of the pre–emphasized speech. Although the software program provides initial estimates of the formant frequencies and bandwidths, the initial estimates are not always accurate values for a wide variety of speech data. The filter parameters can be controlled pitch synchronously as well as pitch asynchronously.

Using this tool, accurate values of formant frequencies and bandwidth along with the glottal excitation signal can be obtained. The estimates of the formant track can be saved for further research, such as speech synthesis, coding, and compression, etc. Theoretically, the spectrum of the differentiated glottal waveform should be white except for the spectral tilt due to the glottal effect to the speech spectrum. Therefore, the goal of the manual inverse filtering procedure is to adjust the filter coefficients of the inverse filter in order that the glottal volume velocity has a maximally flat spectrum. After some trial–and–error, the filter parameters can be fine–tuned manually so that the filter output gives a white spectrum, which means that the vocal tract transfer function is completely cancelled out by the inverse filter.

For example, a speech waveform and its glottal waveform using an automatic inverse filtering algorithm and its FFT spectrum are shown in Figure 3–9, Figure 3–10, and Figure 3–11. Initial estimates of the first five formant frequencies and bandwidths are shown in the control box as in Figure 3–9. The control box provides the researcher a tool to adjust the transfer function of the inverse filter by changing the formant frequencies and bandwidths manually. The speech waveform, residue signal, differentiated glottal waveform, and glottal waveform are plotted in Figure 3–10 (a) through (d), respectively. Figure 3–11 is the FFT spectrum of the preemphasized speech, transfer function of the inverse filter, and spectrum of the differentiated glottal waveform. The automatic formant estimation algorithm can detect the first two formants well. But the algorithm missed the third formant, causing a high frequency component in the differentiated glottal waveform as in Figure 3–10 (c). This is a typical example of the mismatch of the formant frequencies that can be observed from an automatic glottal inverse filtering algorithm. Due to the formant mismatch, the differentiated glottal waveform does not show distinct glottal closed and open phases.

Figure 3–12 and Figure 3–13 are the results after manual adjustment. Accurate values of the formant frequency and bandwidth are shown in Figure 3–12. Figure 3–13 (c) shows clear glottal open and closed phases. In addition, the high frequency component in the glottal waveform before the manual adjustment (Figure 3–10 (c)) has disappeared. The spectrum of the glottal waveform (or differentiated glottal waveform) is much flatter than before the manual adjustment. It can be seen from Figure 3–14 that the transfer function of the inverse filter matches the spectrum of the speech much better, thus, producing a glottal waveform whose spectrum is closer to white.

The manual inverse filtering in the above example was performed pitch asynchronously. A pitch synchronous manual inverse filtering is also available. In the pitch synchronous method, the analysis is performed in two phases. In the first phase, the pitch information is obtained by identifying glottal closure instance. The sharp peaks in the residue signal filtered by a zero–phase lowpass filter was used to identify the glottal closure instant (Hu, 1993; Oppenheim and Willsky, 1983). In the second phase, the speech sample of one pitch period is inverse filtered using the vocal tract transfer function



Pitch Asynchronous MGIF control box

Filter Gain. [dB] : 200

1st fmt freq.[Hz] : 576

bwd. [Hz] : 67

2nd fmt freq.[Hz] : 1614

bwd. [Hz] : 88

3rd fmt freq.[Hz] : 2535

bwd. [Hz] : 260

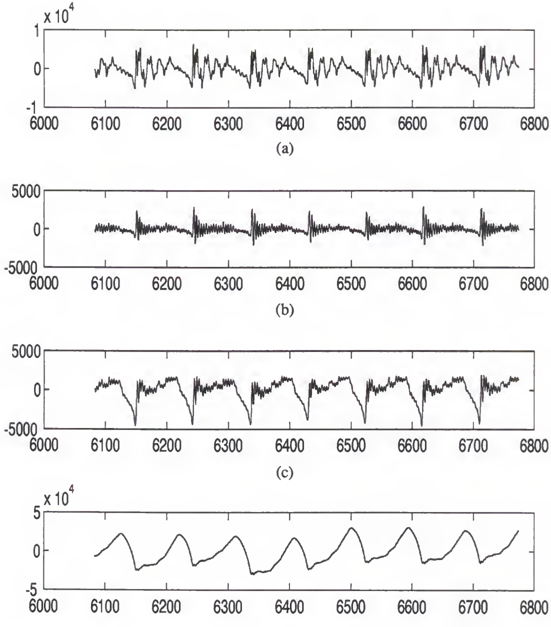
4th fmt freq.[Hz] : 3263

bwd. [Hz] : 419

5th fmt freq.[Hz] : 3453

bwd. [Hz] : 526

Figure 3–9. A control window for the manual inverse filtering before manual adjustment.



(a)

(b)

(c)

(d)

Figure 3–10. Output waveforms of the inverse filter before manual adjustment : (a) Speech waveform (b) Residue signal (c) Differentiated glottal waveform (d) Glottal waveform.

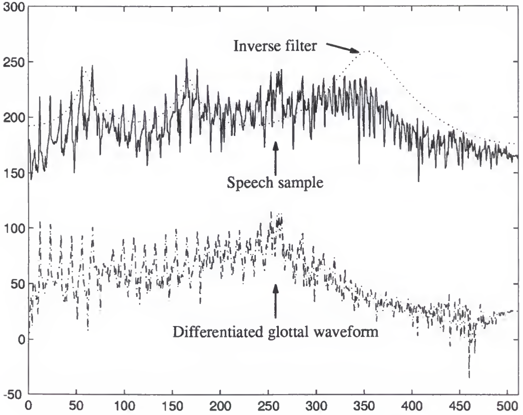
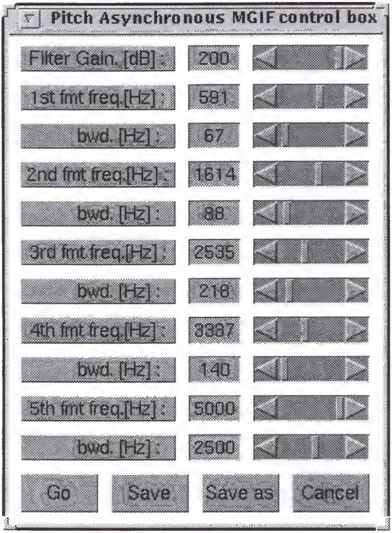


Figure 3–11. The FFT spectrum of the speech sample, inverse filter, and differentiated glottal waveform before manual adjustment.



Filter Gain. [dB] : 200

1st fmt freq.[Hz] : 581

bwd. [Hz] : 67

2nd fmt freq.[Hz] : 1614

bwd. [Hz] : 88

3rd fmt.freq.[Hz] : 2535

bwd. [Hz] : 218

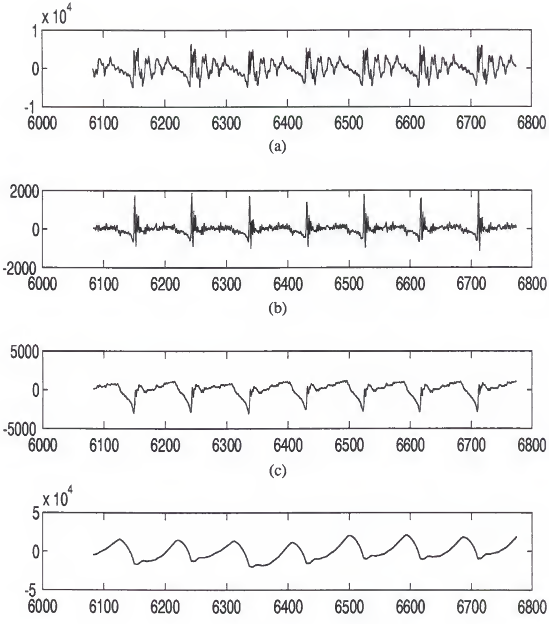
4th fmt.freq.[Hz] : 3387

bwd. [Hz] : 140

5th fmt freq.[Hz] : 5000

bwd. [Hz] : 2500

Figure 3–12. A control window for the manual inverse filtering after manual adjustment.



(a)

(b)

(c)

(d)

Figure 3–13. Output waveforms of the inverse filter after manual adjustment : (a) Speech waveform (b) Residue signal (c) Differentiated glottal waveform (d) Glottal waveform.

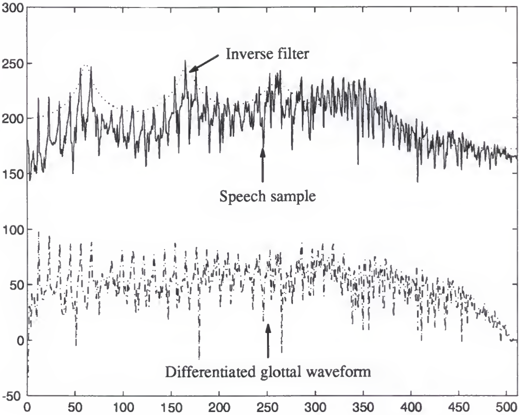


Figure 3–14. The FFT spectrum of the speech sample, inverse filter, and differentiated glottal waveform after manual adjustment.

simulated by the manual control box. The pitch synchronous manual inverse filtering procedure can trace the glottal and vocal tract variance accurately.

Although the manual inverse filtering can be applied to a wide variety of speech data (vocal quality and pitch period), there are several disadvantages. One disadvantage of the manual inverse filtering algorithm is the procedure is highly dependent on the researcher's subjective criteria. Therefore, the inverse filtering output is not always consistent. The other one is that the manual inverse filtering is a time consuming job compared to the automatic algorithms. Nevertheless, the manual inverse filtering algorithm can provide a reliable tool for research of pathological voices and female voices for which the automatic inverse filtering algorithms do not provide good results.

3.3.2 Selective Inverse Filtering

The manual inverse filtering algorithm can also be applied to selective inverse filtering. Figure 3–15 and Figure 3–16 are results of selective inverse filtering in which only the first formant is filtered out. The second and third formants are intentionally mismatched in order to see the effect of these two formant frequencies and bandwidths on the glottal waveform. In this case, the differentiated glottal waveform has a high frequency component, mainly from the remaining second formant. However, the glottal waveform, Figure 3–15 (d), which is obtained by integrating the differentiated glottal waveform, Figure 3–15 (c), show a portion of the glottal open and closed phases.

Figure 3–17 is the result of selective inverse filtering when only the second formant is filtered out. Therefore, the component of the first and third formants are still included in the inverse filtered output spectrum as in Figure 3–18. Figure 3–19 and Figure 3–20 are results of selective inverse filtering in which only the third formant is filtered out. From this experiment, we can conclude that the first two formants are critical ones to achieve a reliable glottal waveform by glottal inverse filtering.

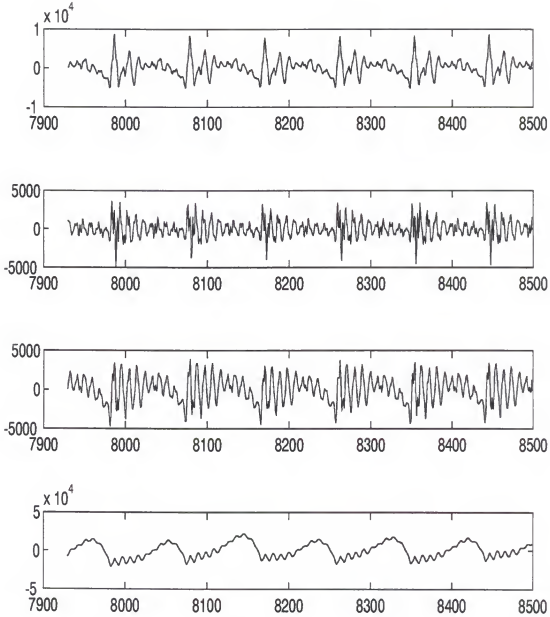


Figure 3–15. Output waveforms of the selective inverse filtering of the first formant :
(a) Speech waveform (b) Residue signal (c) Differentiated glottal waveform (d) Glottal waveform.

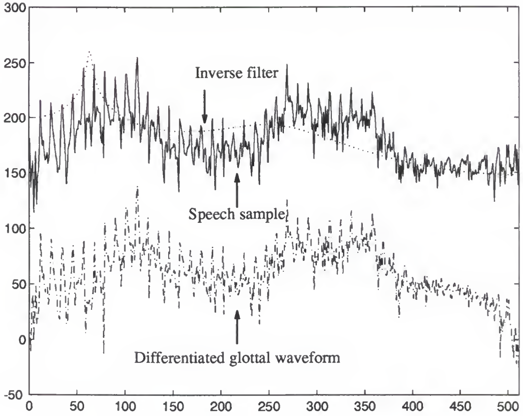


Figure 3–16. The FFT spectrum of the speech sample, inverse filter, and differentiated glottal waveform for selective filtering of the first formant.

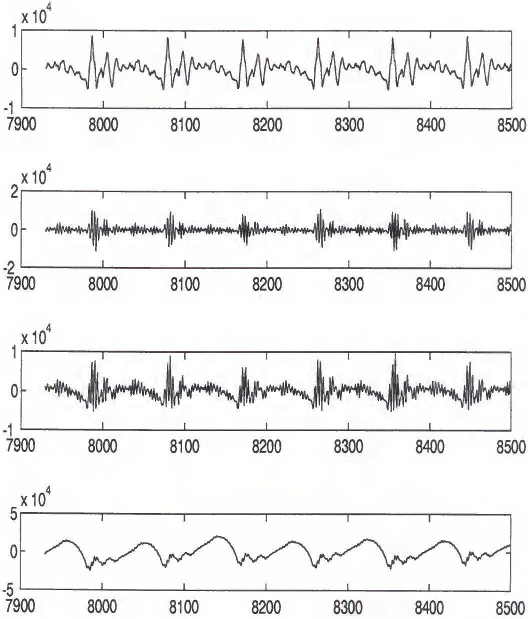


Figure 3–17. Output waveforms of the selective inverse filtering of the second formant :
(a) Speech waveform (b) Residue signal (c) Differentiated glottal waveform (d) Glottal waveform.

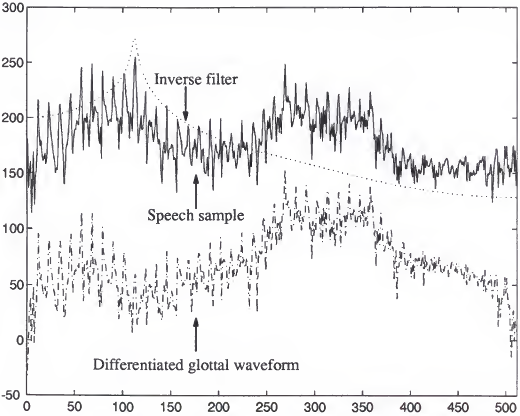


Figure 3–18. The FFT spectrum of the speech sample, inverse filter, and differentiated glottal waveform for selective filtering of the first formant.

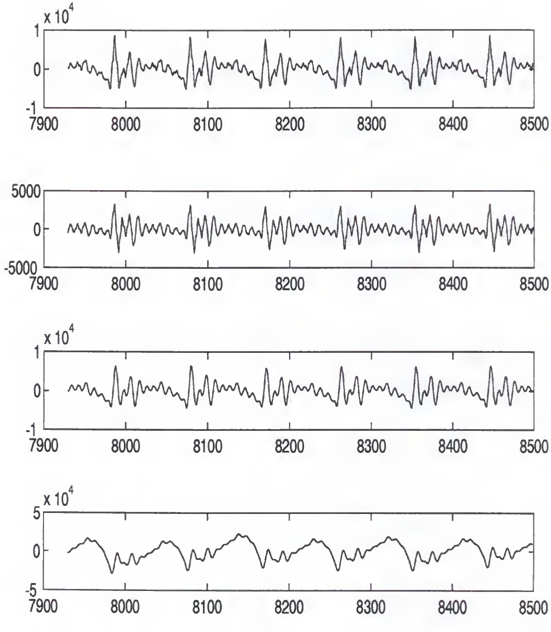


Figure 3–19. Output waveforms of the selective inverse filtering of the third formant :
(a) Speech waveform (b) Residue signal (c) Differentiated glottal waveform (d) Glottal waveform.

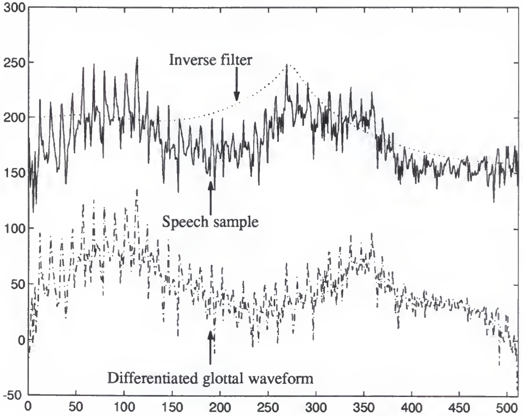


Figure 3–20. The FFT spectrum of the speech sample, inverse filter, and differentiated glottal waveform for selective filtering of the third formant.

In this chapter, the research design is explained and some analysis tools that are used in the research are described. The next chapter will show the analysis results for unvoiced fricatives. The unvoiced fricative production model parameters estimated from the analysis phase will be used to generate synthetic fricatives using an articulatory speech synthesizer.

CHAPTER 4
EXPERIMENTAL RESULTS

4.1 Speech Data Base

The data base used in this research consisted of recordings of sustained fricatives from several speakers. The subjects served in this research are normal (modal) subjects who had no history of vocal disorders or laryngeal pathology. All subjects were male. The data are summarized in Table 4–1.

All data were recorded inside an Industrial Acoustics Company (IAC) single–wall sound room. The speech signals collected have a high signal–to–noise ratio (SNR) and noise can be effectively ignored. A microphone (either an Electro–Voice RE–10 dynamic cardioid microphone or a Bruel & Kjaer (B&K) model 4113 condenser microphone, depending on the task recorded) was located at a fixed distance of 6 inches from the speaker's lips.

Before digitization, the speech was bandlimited to 5 kHz by anti–aliasing, passive, elliptic filters with a minimum stopband attenuation of –55 dB and a passband ripple of ±0.2 dB. The signal was amplified by a Digital Sound Corporation DSC–240 audio control console. The speech was directly digitized at a sampling frequency of 10 kHz per channel by a Digital Sound Corporation DSC–200 A/D and D/A system with 16–bit precision.

4.2 Microphone Characteristics

When the speech signal is used for glottal source estimation, the recording device should have a good low–frequency response. The reason for this is that the glottal source

Table 4–1. The data base* for unvoiced speech analysis.

Subject	Sex/Age	Phonation type	Data file	Contents	Microphone
DMH	M/37	Modal voice	dmhn015	/h/ in hat	EV**
			dmhn016	/f/ in fix	EV**
			dmhn017	/θ/ in thick	EV**
			dmhn018	/s/ in sat	EV**
			dmhn019	/ʃ/ in ship	EV**
			dmhn020	/v/ in van	EV**
			dmhn021	/ð/ in this	EV**
			dmhn022	/z/ in zoo	EV**
			dmhn023	/ʒ/ in azure	EV**

*Natural fricative speech with normal vocal quality

**Microphone Type: E = Electro–Voice RE–10 waveform, which is to be estimated, has its major energy components at low frequencies (dc to 1 kHz). Of the two microphones used to measure the sound pressure waveforms, the B&K 4133 condenser microphone has the best low–frequency response. Its amplitude response is within ±1 dB down to 20 Hz, and its phase response is linear. The –3 dB low–frequency cut–off is approximately 10 Hz. Because of this good low–frequency characteristic, the B&K 4133 condenser microphone is also sensitive to low–frequency breath and ambient noise, which may cause problems in speech analysis. Therefore, the Electro–Voice RE–10 microphone was used to collect most of the speech data.

The Electro–Voice RE–10 microphone has a good frequency response at frequencies above 50 Hz, but attenuates the low–frequency components below 50 Hz. When compared to the B&K 4133 condenser microphone, the obvious drawback of the Electro–Voice RE–10 microphone is the lack of good low–frequency response. Thus, for the purpose of glottal inverse filtering, for which a good low–frequency response is required, speech data collected by using the Electro–Voice RE–10 microphone had to be corrected to compensate for the low–frequency distortion, based upon the characteristics of the B&K 4133 condenser microphone [Wong, 1991].

4.3 Noise Source Analysis

4.3.1 Assumptions

The experiments for this research are based on several assumptions regarding the generation of unvoiced speech.

$\boxed{1}$ When a turbulence noise is generated at constrictions that are located either at the glottis or in the vocal tract, it is assumed that only one constriction is dominant and, thus, the effect of the other constriction can be neglected.

For aspiration, the glottal constriction is a major obstacle of the air flow, while for fricatives a constriction in the vocal tract formed (supraglottal constriction) by the tongue tip or tongue body is a main factor generating a turbulence noise.

$\boxed{2}$ For some fricative sounds generated from a supraglottal constriction, the effect of back cavity resonance to the speech spectrum can be neglected if the constriction length is narrow and long (Heinz and Stevens, 1961). This assumption will be used without proof in the analysis phase, and will be verified later in the synthesis phase. Using an articulatory synthesizer, fricative sounds will be generated with a variable constriction length and the effect of the length will be evaluated.

$\boxed{3}$ It has been confirmed by previous research that the estimates of the effective second formant frequency using the PLP algorithm is close to the front cavity resonance frequency for voiced sounds. In this research, the assumption is that the PLP algorithm is appropriate for unvoiced sounds.

$\boxed{4}$ The front cavity transfer function can be modeled by a 5th order all–pole model.

4.3.2 Analysis Procedure

An analysis algorithm based on the auditory model was implemented as explained in the previous chapter. The overall analysis algorithm is summarized in Figure 4–1. The main purpose of the analysis program is to estimate the front cavity resonance characteristics, which might be an important factor for generation of unvoiced fricatives sound. The front cavity resonance corresponds to the quarter–wavelength resonance frequency of the front cavity (Kuhn, 1975). The analysis results are used to synthesize unvoiced sounds during the synthesis phase, which will be described later in this chapter.

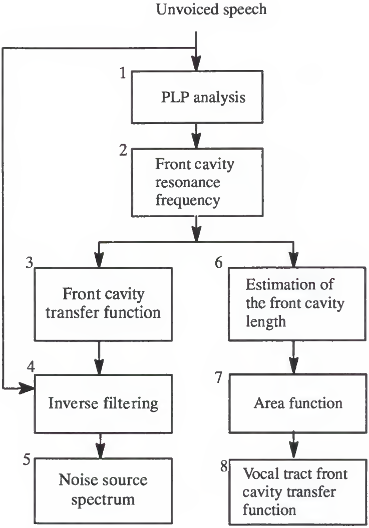

Unvoiced speech
1
PLP analysis
2
Front cavity resonance frequency
3
Front cavity transfer function
4
Inverse filtering
5
Noise source spectrum
6
Estimation of the front cavity length
7
Area function
8
Vocal tract front cavity transfer function

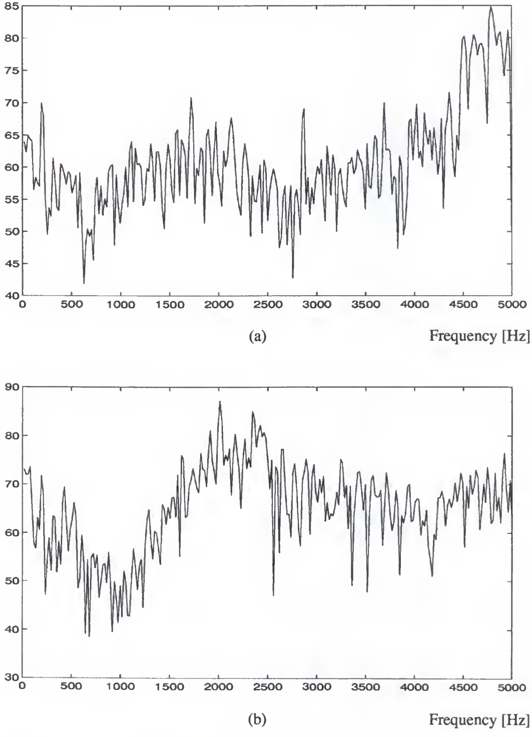
Figure 4–1. Overall analysis procedure.

Before the speech is analyzed, the speech signal is segmented into frames of 256 points, which corresponds to 25.6 msec, and windowed using a 256 points Hamming window. First, the spectrum of the speech segment is transformed into the Bark scale and after that, all the auditory modeling processes, such as equal loudness pre–emphasis, intensity–loudness pre–emphasis, are performed in the Bark scale. Finally, the auditory spectrum is represented using an all–pole model. It has been shown in speech recognition experiments that a 5th order PLP algorithm has a performance equivalent to a 13th order conventional LP model (Hermansky, 1990). In this research, we adopt an all–pole model of 5th order and, therefore, five poles are obtained from the PLP algorithm. The pole frequency and bandwidth are transformed back to the linear scale (Hz) from the Bark scale. Poles that have large bandwidth are eliminated. A pole frequency that corresponds to a minimum bandwidth is chosen as a major spectral peak frequency. The frequency is our estimate of the vocal tract front cavity resonance frequency. Once the front cavity resonance frequency is estimated, one can estimate the functional length of the front cavity, which is the length from the lips to a supraglottal constriction, using the quarter–wavelength formula, Equation 3–6.

4.3.3 Experiments

In this section, some analysis results will be explained. Analysis data used are sustained fricatives of /s/, /ʃ/, /f/, and /h/. The FFT spectra for each of the four fricative classes are shown in Figure 4–2. The spectra are computed using a 512 point FFT for Hamming windowed speech segments.

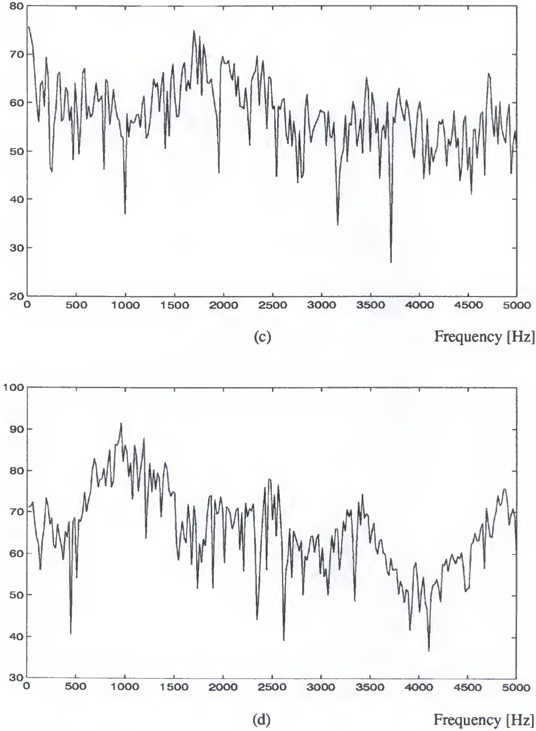
The speech segments are extracted manually in the most stable part of each sound. Since fricative sounds are generated by a stable configuration of the articulators, no information about the dynamics is needed. The spectrum of /s/ shows one major spectral peak near 5000 Hz, while for /ʃ/, there is a spectral peak at about 1900 Hz. The spectrum of /f/ does not have a prominent peak and the spectrum of /h/ has a peak at a low



(a)

(b)

Figure 4–2. Example of measured spectra for (a) /s/, (b) /ʃ/, (c) /f/, and (d) /h/.



(c)

(d)

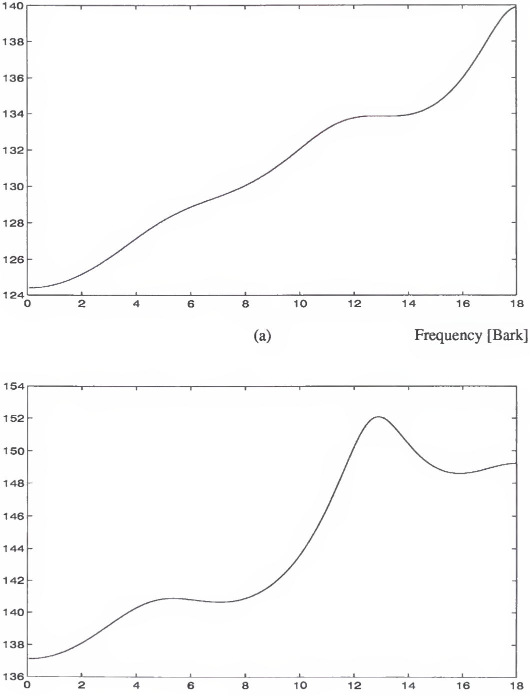
Figure 4–2. Continued.

frequency below 1000 Hz. There is only a small difference in the frequency region below 1000 Hz for the fricatives /s/, /ʃ/, and /f/.

4.3.3.1 Experiment 1 – Estimation of the front cavity resonance

In order to find which spectral peaks are perceptually important, the PLP algorithm of 5th order was applied to the same speech segments. The PLP spectra are depicted in Figure 4–3. We presume that the peak frequency corresponds to the resonance of the front cavity as per previous research. Since the abscissa is in the Bark scale, one can transform the peak frequency in Bark to the linear scale as summarized in Table 4–2. The PLP spectrum of /s/ shows one major spectral peak near 5000 Hz with a bandwidth of 279.7 Hz. The estimates for a major spectral peak are almost the same as that from the FFT spectrum for /s/ sound. However, the spectrum of /ʃ/ has a spectral peak at 2151.9 Hz, which is a little higher than the estimate from the FFT spectrum. The bandwidth is about 240.1 Hz. The spectrum of /f/ does have a prominent peak in the PLP spectrum. However, its bandwidth is too wide to be considered as a formant. The fact that there is no front and back cavity in generating /f/ sound explains why the PLP algorithm does not give a front cavity resonance estimate with a reasonable bandwidth. The PLP spectrum of /h/ has a peak at low frequency below 1000 Hz with a bandwidth of 342.5 Hz.

We can conclude that the PLP algorithm can detect the front cavity resonance for /s/, /ʃ/, and /h/ sounds. For /f/ sounds, the estimate has too wide a bandwidth. The front cavity resonance can be estimated by choosing a pole frequency, which has a bandwidth less than a threshold value. From this experiment, a threshold for the bandwidth of about 400 Hz seems a reasonable value. If there are no poles that meet the bandwidth requirement, it is assumed that the place of articulation is at the mouth opening, as in the example of /f/, and /θ/.



(a)

(b)

Figure 4–3. Example of the PLP spectra for (a) /s/, (b) /ʃ/, (c) /f/, and (d) /h/.

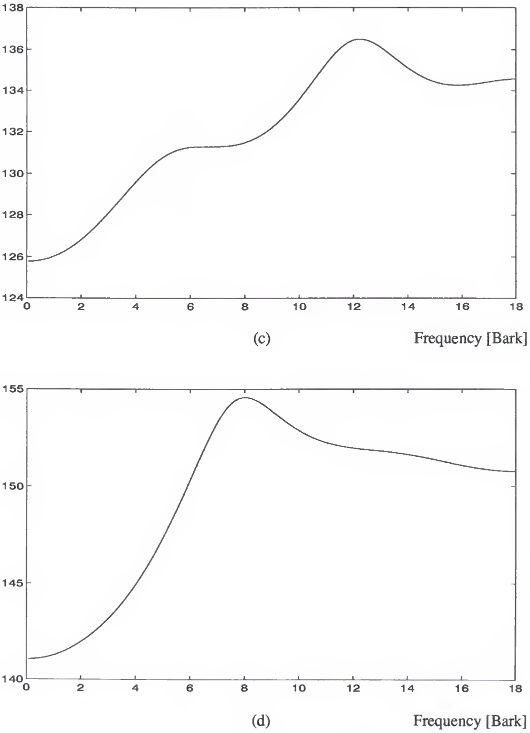


Figure 4–3. Continued.

Table 4–2. Table of PLP spectral peaks frequencies and bandwidth in both Bark scale and the linear scale.

Data	Spectral Peak	Peak frequency		Peak bandwidth	
		[Bark]	[Hz]	[Bark]	[Hz]
/s/ in sat	F_1'	10.9	1809.0	5.2	593.8
	F_2'	17.0	5083.0	2.7	279.7
/ʃ/ in ship	F_1'	11.9	2151.9	2.3	240.1
	F_2'	17.0	5083.0	4.7	525.2
/θ/ in thick	F_1'	10.7	1784.4	7.1	889.7
	F_2'	17.0	5083.0	12.1	2197.3
/f/ in fix	F_1'	4.8	535.0	5.0	565.1
	F_2'	11.3	1917.7	3.8	407.9
/h/ in hat	F_1'	7.1	887.7	3.3	342.5
	F_2'	12.6	2399.5	8.4	1142.3

4.3.3.2 Experiment 2 – Estimation of the front cavity length for sibilant fricatives

As explained in the previous chapter, the front cavity resonance frequency can be regarded as the quarter–wave resonance. This experiment is to prove the assumption that the front cavity resonance frequency can be estimated by the frequency of the main spectral peak in the PLP spectrum. Using Equation 3–6, and the front cavity resonance estimates summarized in Table 4–2, the vocal tract front cavity length can be calculated. Table 4–3 is the calculated value of the front cavity length. The speed of sound is approximated as 353 m/sec (for 35°C) in the calculation. For sibilant fricative sounds, such as /s/ and /ʃ/, the estimated length are quite reasonable. On the other hand, there are no front cavity and back cavity for /θ/ and /f/ sounds, and, therefore, the analysis algorithm could not find a front cavity resonance frequency that has a bandwidth of less than 400 Hz. In this case, the place of articulation is at the mouth opening, and thus, the length of effective front cavity is zero. Finally, the effective front cavity length of 9.94

Table 4–3. The estimates of the effective vocal tract front cavity length.

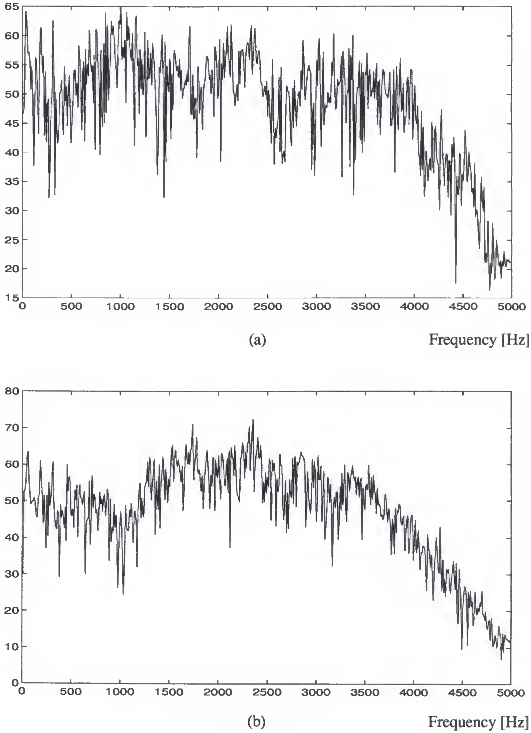
	/s/ in sat	/ʃ/ in ship	/θ/ in thin	/f/ in fix	/h/ in hat
Front cavity length (cm)	1.74	4.1	0	0	9.94

cm for aspiration noise /h/ is shorter than we expected. The constriction is located at the glottis for aspiration noise and the average vocal tract length of male speaker is about 17cm. This error might be due to the low velocity air flow with less constricted vocal tract in generating aspiration noise. From this experiment, we can conclude that the front cavity resonance frequency moves upward in frequency as the distance from the consonant constriction to the lips decreases and the analysis algorithm based on the PLP algorithm can provide a reliable estimate of front cavity resonance. When the constriction is even more anterior, there is no front cavity and the analysis algorithm does not produce any reasonable estimate of the front cavity resonance, in which case we can assume that the constriction is at the mouth opening.

4.3.3.3 Experiment 3 – Estimation of noise source spectrum

The turbulence noise source spectrum is estimated by inverse filtering the unvoiced speech signal. The inverse filter for each fricative, which is the reciprocal of the vocal tract front cavity transfer function, is constructed using the pole frequencies and bandwidths shown is Table 4–2. Air flow at the noise source is obtained by filtering the speech signal with the front cavity inverse filter, and the spectrum of the turbulence noise source spectrum is calculated. The estimates of turbulence noise source spectrum for /s/, /ʃ/, /f/, and /h/ are shown in Figure 4–4 .

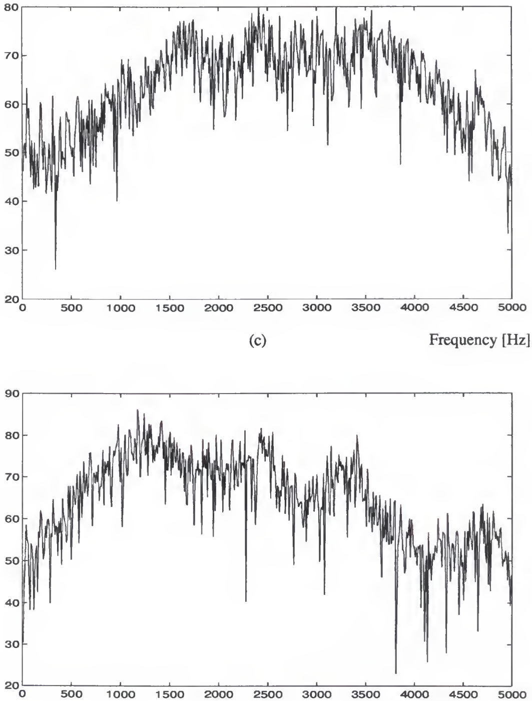
In general, the source spectrum has a peak with wide bandwidth. The center frequency of the peak varies according to the place of articulation, i.e., location of the constriction. The center frequency is in the range of 2000–2500 Hz for most types of frication noise such as /s/, /ʃ/, and /f/, while the center frequency for aspiration noise /h/ is about 1000 Hz. This observation is in accordance with previous simulation results (Stevens, 1971). Table 4–4 summarizes the center frequency of turbulence noise source spectrum for different kinds of fricative sounds. It can be observed that as the place of constriction becomes close to the lips, the center frequency goes higher.



(a) Frequency [Hz]

(b) Frequency [Hz]

Figure 4–4. Estimates of the turbulence noise source spectra for (a) /s/, (b) /ʃ/, (c) /f/, and (d) /h/.



(c)

(d)

Figure 4–2. Continued.

Table 4–4. The estimates of center frequency of turbulence noise sources for unvoiced speech.

	/s/ in sat	/ʃ/ in ship	/f/ in fix	/h/ in hat
Center frequency (Hz)	2300	2300	2800	1100

4.3.4 Discussion

The analysis method adopted in this research is based on the psychoacoustics of human hearing. Research in linguistics and psychoacoustics suggest that the human auditory system tends to simplify the formant information by extracting the effective second formant. The effective second formant can be regarded as a combination of formants. The concept of effective second formant is applied here to the analysis of unvoiced speech. The spectral peak in the auditory spectrum is regarded as the front cavity resonance and used to estimated the spectrum of noise source and the effective length of the front cavity. In the next section, we will verify the analysis method based on the auditory model by synthesizing unvoiced sounds according to the parameters obtained in analysis phase.

4.4 Synthesis of Unvoiced Fricatives

In order to verify the validity of the parameters for the unvoiced speech generation model, unvoiced sounds are generated using an articulatory synthesizer developed by Hsieh (1994). In this section, the articulatory synthesizer will be explained briefly and a new unvoiced fricative speech production model will be introduced.

4.4.1 Articulatory Synthesizer

Basically, there are three approaches used in articulatory speech synthesis. The wave digital filter approach (Fettweis and Meerkötter, 1975; Lawson and Mirzai, 1990) that extended the Kelly-Lochbaum model (1962). The second approach uses a hybrid time-frequency domain method, which models the highly nonlinear glottal characteristics in the time domain and the linear tract with frequency-dependent losses and wall vibration characteristics in the frequency domain (Allen and Strong, 1985; Sondhi and Schroeter, 1986, 1987). The third approach is to model the human vocal system as a large set of, linear or nonlinear, difference equations to be solved in each sampling interval to give samples of the pressure and volume velocity at each point in the transmission-line circuit (Flanagan and Cherry, 1968; Flanagan and Landgraf, 1968; Flanagan and Ishizaka, 1976; Flanagan et al., 1975, 1980). The values of pressure and volume velocity at one time instant are used to determine the losses for the next time interval. This approach has been referred to as the time-domain approach (Sondhi and Schroeter, 1987).

In the time-domain approach, a very high sampling rate is usually required to avoid frequency-warping distortion (Wakita and Fant, 1978). Several advantages have made the time-domain approach popular, although its computation is cumbersome. These advantages are that the aerodynamic interaction is inherently included, the pressure and volume velocity at any point can be computed, and the dynamic articulatory gestures can be obtained when combined with the articulatory model. In our study, the time-domain approach is used for articulatory speech synthesis (Hsieh, 1994).

4.4.2 Unvoiced Fricative Production Model

According to the above analysis results, a new unvoiced fricative production model can be drawn as in Figure 4–5. The noise generator is implemented based on the model proposed by Sondhi and Schroeter (1986, 1987). The noise source model defines the characteristics of the noise source as a function of the airflow through the constriction and of the constriction cross-sectional area A_c. This model allows the user to place the turbulence noise source at the center of, or immediately downstream or upstream from the constriction region, or spatially distributed along the constriction region. The turbulence gain and critical Reynolds number can also be specified. The magnitude spectrum of the noise generated from block 1 in Figure 4–5 is white. The white noise spectrum is shaped by a spectral shaping filter, which simulates the noise source spectrum estimated in the analysis phase.

The source spectral shaping filter is a low–order linear prediction filter. The filter coefficients are obtained by modeling the turbulence noise source spectrum using an all–pole model. Since the turbulence noise spectrum has a wide spectral peak at the center frequency of F_c, the filter can be represented by a low–order (e.g., 3rd order) linear prediction filter (Stevens, 1971).

The articulatory synthesizer is based on the time–domain approach suggested by Sondhi and Schroeter (1987) and improved by Hsieh (1994). The input area function is transformed to the equivalent RLC–network. Using the noise source excitation from the spectral shaping filter as a voltage source and by applying Kirchoff's and Ohm's law to the network, the discrete–time acoustic matrix equations are formed. The pressure at the midpoint of each section and volume velocity are calculated as solutions using the elimination procedure and a backward substitution. The synthetic speech is the backward difference between the sum of the volume velocities at the nostrils and lips at the current

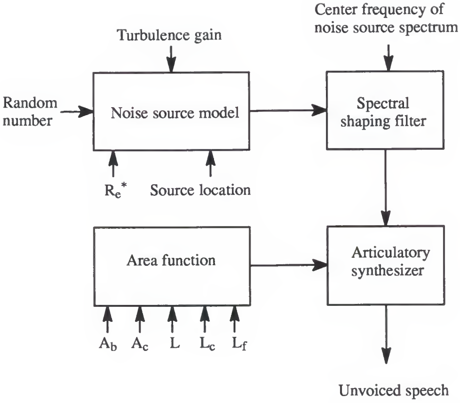


Figure 4–5. A block diagram of unvoiced fricatives generation model. Refer to Figure 3–2 for the values of the input parameters to the area function.

time and the sum of the volume velocities at the nostrils and lips at the previous time instant (Hsieh, 1994)

In this research, the production model is used to generate unvoiced fricatives. Some examples of synthetic speech are explained in the next section along with the results for the experiments.

4.4.3 Unvoiced Fricative Speech Synthesis

Unvoiced fricative sounds are generated using the articulatory synthesizer explained in the previous section. In order to investigate the effects of the front cavity length, supraglottal constriction length, and back cavity shape on synthetic speech sounds, the following experiments are considered.

4.4.3.1 Experiment 4 – Effect of the front cavity length on the synthetic speech

With the noise source parameters fixed, the effect of the front cavity length on the synthetic fricative sound can be determined by comparing the spectra of the synthetic sounds generated by changing the front cavity length. The length of the constriction and the back cavity width are fixed in this experiment. Figure 4–6 shows the area function and the FFT spectra for unvoiced sounds. The cross sectional area for the constriction is set to 0.1 cm^2, and the front cavity length is gradually increased from 0 cm (Figure 4–6 (a)) to 6.4 cm (Figure 4–6 (i)). We can compare the FFT spectrum of each configuration with the spectrum measure from real fricative speech for /s/, /ʃ/, /f/, and /h/ sounds (Figure 4–2). When the front cavity length is zero (Figure 4–6 (a)), which means the place of articulation is at the lips, the spectrum of the synthetic sound is similar to the spectrum of the /f/ sound (Figure 4–2 (c)) showing no spectral tilt. As the front cavity length increases to one and two centimeters (Figure 4–6 (b)), the spectrum becomes close to the spectrum of the /s/ sound (Figure 4–2 (c)), in which the level of the spectral peak

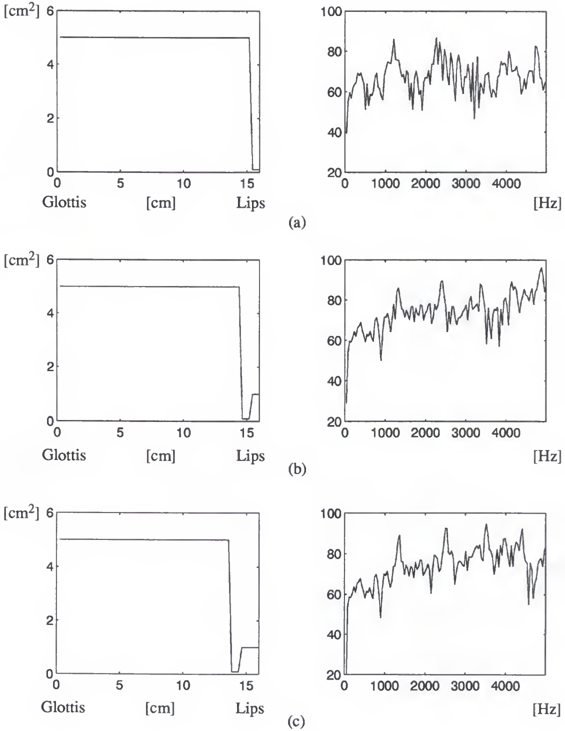


Figure 4–6. Area function and FFT spectrum of synthetic speech.
Only the front cavity length is varied.

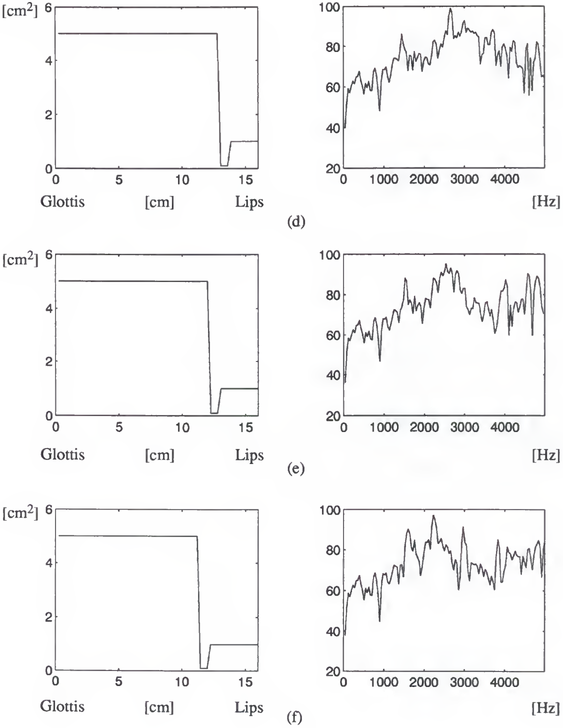


Figure 4–6. Continued.

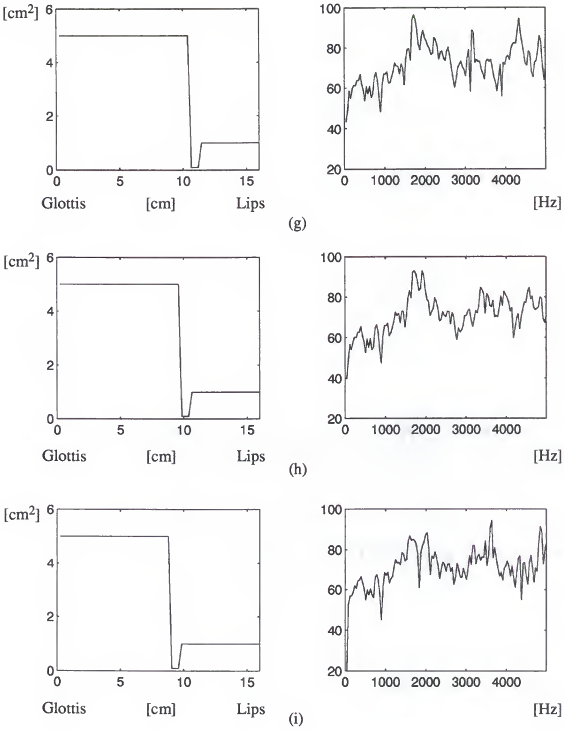


Figure 4–6. Continued.

goes higher as the frequency increases. Thus, the major peak is around 5000 Hz as in the measured spectrum of /s/. When the length of the front cavity is around three to six centimeters (Figure 4–2 (d), through (h)), the spectrum resembles the spectrum of /ʃ/, which is a reasonable front cavity length. The front cavity length for each synthesized sound is reasonable when compared with the known production of the corresponding real speech samples (refer to Table 4–3)

The observations in this experiment are confirmed by informal listening tests. In conclusion, the front cavity length is a critical parameter deciding the place of articulation and, thus, the fricative sounds produced. Also, we have confirmed that the estimated value of the front cavity length is valid.

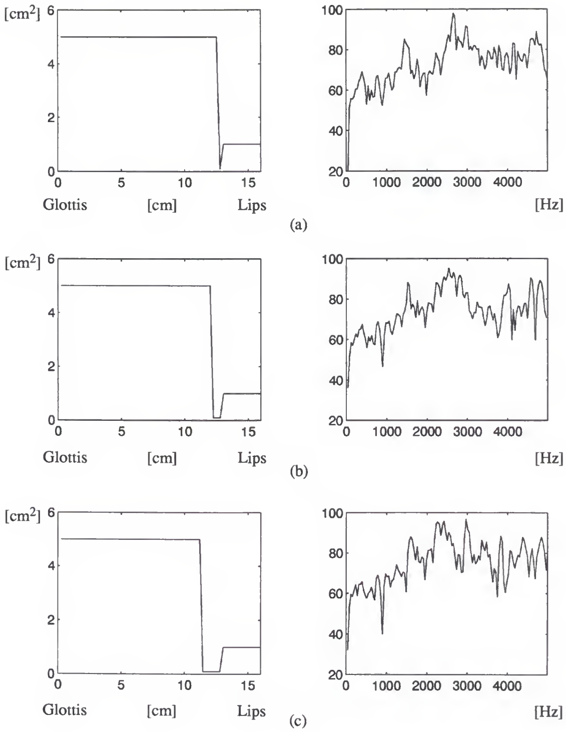
4.4.3.2 Experiment 5 – Effect of the constriction length on the synthetic speech

The effect of the constriction length can be examined from this experiment, where unvoiced speech is generated using a variable constriction length with the front cavity length fixed. Notice three spectral peaks (formants) in Figure 4–7 (a). As the length of constriction increases, the second formant F_2 increases and the third formant F_3 decreases in frequency. However, the first and fourth formants remain fixed.

Informal listening tests indicate that although the FFT spectra are different as shown in Figure 4–7 (a) through (f), the perceptual difference is not significant. We can postulate that the back cavity resonance change does not affect the overall auditory spectrum. In the next experiment, we change the shape of back cavity.

4.4.3.3 Experiment 6 – Effect of the back cavity on the synthetic fricative sounds

This experiment is to confirm that the back cavity resonance does not have much effect on the synthetic fricative sound as long as the supraglottal constriction is sufficiently narrow and long. Given noise source parameters, a front cavity length, and a

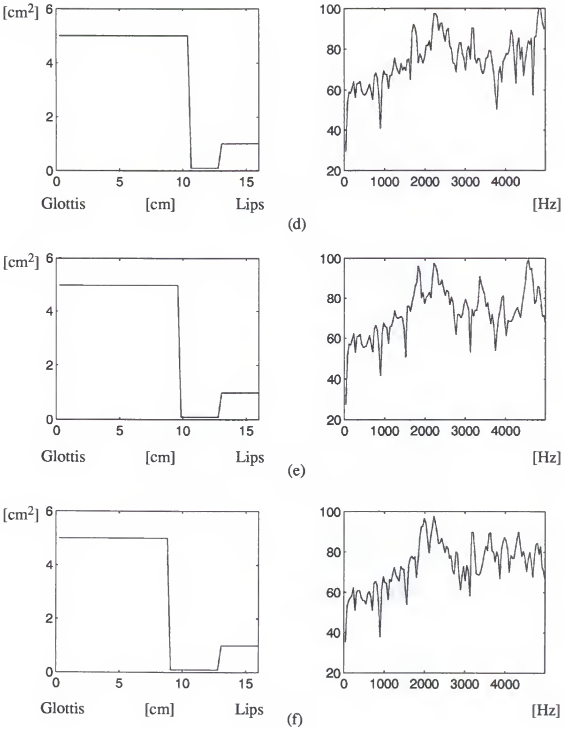


(a)

(b)

(c)

Figure 4–7. Area function and FFT spectrum of synthetic speech.



(d)

(e)

(f)

Figure 4–7. Continued.

constriction length/width, the shape of the area function corresponding to the back cavity is changed as in the left column in Figure 4–8. The front cavity and constriction part of the three area functions are almost the same, i.e., the cross sectional area of the supraglottal constriction is 0.4 cm^2, the front cavity length is 1.9 cm, and the front cavity area is 5 cm^2. These numerical value are based on the measured area function of Figure 4–8 (c) using x–rays.

The synthetic speech generated is close to fricative /ʃ/. Since the vocal tract configuration for /ʃ/ has a distinct front cavity and back cavity, it is easy to observe the effect of the back cavity resonance on the speech spectrum. Synthetic fricatives generated from two different back cavity shapes are compared with a synthetic fricative from the measured area functions using X–rays (Badin, 1991) in Figure 4–8.

As expected. the formant structure of the three spectra are very similar because the back cavity resonance is decoupled from the constriction and the front cavity, thus, having little influence on the overall speech spectrum. The results of this experiments are in agreement with the fricative speech production theory, i.e., the back cavity resonance does not play an important role in generating fricative sounds.

4.5 Discussion

The main point of the experiments in this research is the importance of the front cavity resonance in the production of unvoiced fricative speech. The length of the front cavity can be estimated from the speech signal using the analysis method developed here, which is based on psychoacoustics. In general, a simple vocal tract area function is enough to generate an intelligible unvoiced sound. From the experiments on unvoiced fricatives speech, we can summarize the production of unvoiced speech as follows:

$\boxed{1}$ For sibilant fricatives (/s/ and /ʃ/), the supraglottal constrictions, which are configured by the tongue body and hard plate for /ʃ/ and by the tongue tip

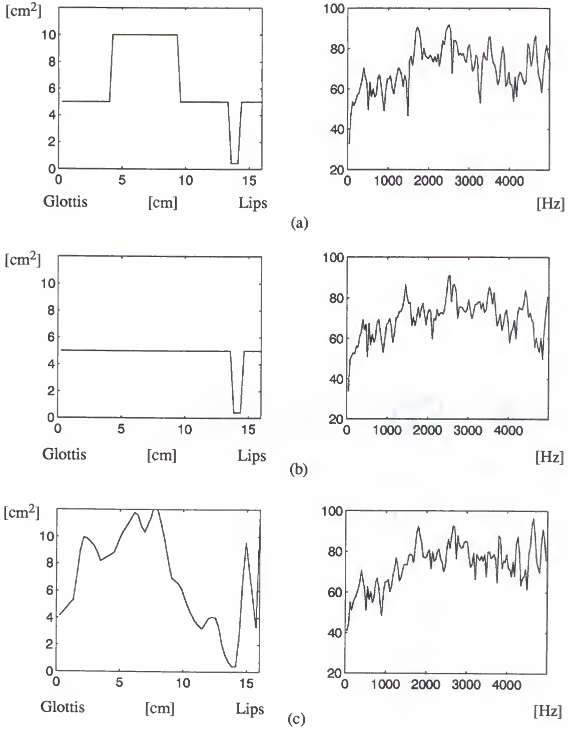


Figure 4–8. Area function and FFT spectrum of synthetic speech.

and tooth ridge for /s/, are narrow and long enough to decouple the back cavity resonance with the front cavity. The length of the front cavity determines which sound is pronounced. And the length can be estimated using the analysis methods developed here, which are based on psychoacoustics.

$\boxed{2}$ The fricative /f/ is generated at the lower lip and upper incisor, and /θ/ is generated by the tongue tip and incisors. The turbulence noise source for these fricatives is located at the mouth opening and, therefore, there is no front or back cavity. The FFT spectrum of speech sound is similar to the spectrum of turbulence noise source.

$\boxed{3}$ For the aspiration sound /h/, the glottal constriction is a major obstacle for the air flow. Since the vocal tract shape for aspiration is already that of the following vowel, the spectral peak from the PLP analysis should be interpreted as the effective second formant of the following vowel.

CHAPTER 5
CONCLUSION AND FUTURE WORK

5.1 Summary

The primary goal of this research is to study unvoiced speech characteristics. Based on the speech production theory and human hearing experiments, the vocal tract front cavity resonance is important to the perception of unvoiced sounds. An algorithm based on the human auditory model (PLP algorithm) is used to estimate the front cavity resonance. The effective length of the vocal tract front cavity can be calculated using the estimate of the front cavity resonance frequency. The turbulence noise source spectrum is also estimated using the front cavity resonance information. The center frequency of the turbulence noise spectrum is used to estimate the location of the noise source. Using an articulatory synthesizer unvoiced speech is synthesized. The articulatory synthesizer consists of a turbulence noise source generator and a vocal tract filter. Effects of the front cavity length, constriction length, and back cavity resonance on the perception of unvoiced speech are studied by informal listening tests. Also, the effects of the spectrum, especially the center frequency, of the turbulence noise source on the output speech are investigated.

5.1.1 Speech Inverse Filtering

If the inverse problem, which is to determine articulatory parameters and vocal tract shape from the speech signal, can be solved, it can be applied to many speech related areas such as speech coding, data compression, and speech recognition. It also can be used in speech production research as well. From previous research, the analysis–by–synthesis method produced a desirable solution to the inverse problem for voiced speech. Although the solution to this problem was often not unique, an optimal solution could be found by applying some minimization criteria, such as minimal muscle work (Sorokin, 1994).

The analysis of voiced speech has been successful using spectral estimation algorithms such as the linear prediction algorithm, and the weighted recursive least square (WRLS) algorithm (Ting, 1989; Lee, 1992), etc. The acoustic information from the spectral estimation algorithms, i.e., formants, is sufficient to obtain the articulatory parameters using a conventional optimization procedure (Hsieh, 1994, Sorokin, 1994)

On the other hand, the speech inverse problem for unvoiced speech is more complicate than for voiced speech. As Sorokin (1994) pointed out, there has been much less success in solving the inverse problem for unvoiced fricatives. The reason is that an adequate analysis algorithm for unvoiced speech has not been found, or that there is little data for unvoiced speech production. The duration of unvoiced speech is short and, above all, the spectral characteristics are transient over a short time interval, especially for affricates and stops. In addition, stop sounds and aspiration are highly affected by the following vowel, i.e. by coarticulation. Considering that a successful inverse solution for both voiced and unvoiced sounds will open new opportunities in various speech applications, research on unvoiced sounds deserves more attention. In this context, this research is focused on unvoiced speech production, noise source modeling, and calculating the vocal tract area function for generating unvoiced sounds.

5.1.2 Human Auditory Model and Speech Analysis

Research on the human auditory system has suggested many theories that can be directly applied to speech compression and speech recognition. By modeling the masking effect and nonuniform sensitivity level of hearing, auditory data, which the human auditory system cannot perceive, can be reduced without sacrificing speech quality. On the other hand, speaker dependent factors can be separated from the speech signal, improving the performance of speech recognizers (Hermansky, 1988; Ghitza, 1994).

Recently, ideas from the human auditory model have been applied to speech analysis research (Qi, 1992; Kewley–Port and Luce, 1984; Kurowski and Blumstein, 1987). Basic concepts and terminologies that are applied to speech analysis research, including the PLP algorithm, are reviewed in Chapter 2. The PLP algorithm is used to estimate the vocal tract front cavity resonance, which is an important factor in the perception of unvoiced as well as voiced speech. The front cavity resonance is believed to carry the phonetic information of the speech signal and the PLP algorithm is an effective method to estimate the front cavity resonance frequency from the speech signal.

5.1.3 Fricative Production Model

Using the front cavity resonance and turbulence noise source spectrum, a new unvoiced fricative speech production model is proposed. The model consists of a noise source generator with variable center frequency and a vocal tract area function that emphasizes the vocal tract constriction and the front cavity. The back cavity has little effect on the speech spectrum for unvoiced fricative sounds. This model can be easily adopted to a voiced fricative generation model by replacing the noise source generator with a quasi–periodic pulse generator.

5.2 Conclusion and Contributions

In this research, a new fricative generation model and an analysis algorithm based on the human auditory model are developed. The model parameters can be obtained from the proposed analysis algorithm. As in the case of voiced speech, the front cavity resonance is important in perception of fricative sounds. Therefore, we can simplify the back cavity part of the vocal tract area function for fricatives. The front cavity length can be estimated from the speech signal. The location of the turbulence noise source also can be estimated from the noise source spectrum.

The results of this research can be applied to the inverse problem for unvoiced speech. To solve an optimization problem for unvoiced speech using the analysis–by–synthesis method based on an articulatory model, it is often required to start the optimization with the proper initial values for the articulatory parameters to obtain stable articulatory solutions (Sorokin, 1994; Schroeter and Sondhi, 1994). The front cavity length estimated in this research can help select the proper initial articulatory position for the optimization problem.

5.3 Future Work

5.3.1 Analysis of Affricates and Stop Consonants

Although, the proposed analysis method provides new knowledge about unvoiced fricatives production, the dynamics of the unvoiced consonants are to be further investigated. Especially unvoiced sound groups with transient spectral characteristics, such as affricates and stop consonants, and the dynamics of the vocal tract front cavity resonance should be examined. In addition, the nonlinear coarticulation effect between unvoiced consonants and following vowels can be studied using the auditory based analysis method explained in this research, i.e., in terms of the effective second formant, or vocal tract front cavity resonance.

5.3.2 Speech Inverse Filtering

Starting the estimates of the vocal tract area function provided from this research, performance of the inverse filtering for unvoiced speech can be improved. The optimization processing time can be reduced using an adequate starting point for the area function. The estimated articulatory position should be more reliable.

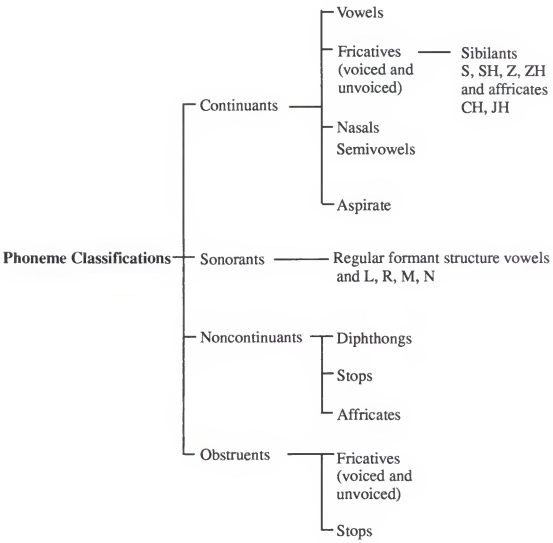
5.3.3 Speech Coding based on Articulatory Production Model

The result of the speech inverse filtering can be applied to a new approach to the very–low–rate speech coding technique. This proposal is described in more detail in Appendix B. The proposal focuses on the fact that the articulators vary slowly with time, and, therefore, can be utilized to segment speech for the purpose of speech coding. The proposed technique is to segment speech according to phonetically related intervals that are synchronized with articulatory movement. If the segments are longer than those presently used with the pitch asynchronous method, then we may be able to code speech according to articulatory movement and achieve a low bandwidth, high quality speech coding scheme. The articulatory position and its movement could be estimated using the acoustic information obtained by glottal inverse filtering and an acoustic–to–articulatory transformation technique.

Recently, this topic has become of interest to researchers because the quality of known coding algorithms at low bit rates is unacceptably poor and the computing power required to implement the complex algorithms based on the articulatory production model is now available.

APPENDIX A
PHONEME CLASSIFICATIONS AND SYMBOL TABLE

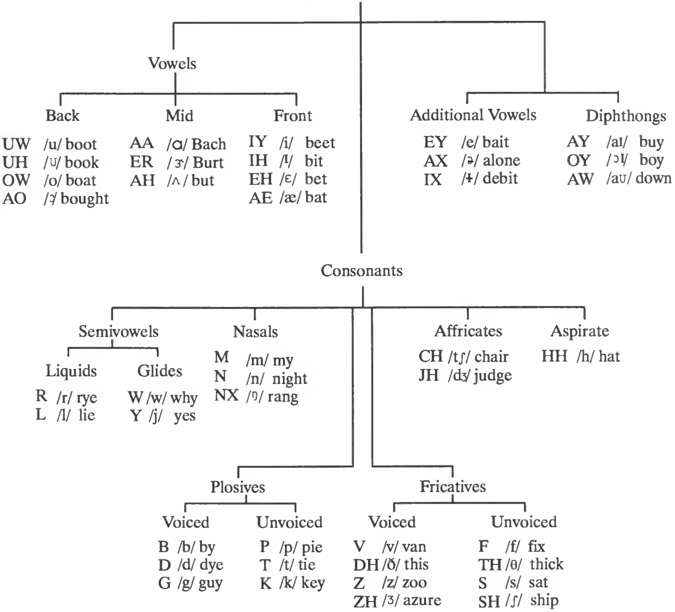
A.1 Phoneme Classification based on Phonation Method



Phoneme Classifications
Continuants
Vowels
Fricatives (voiced and unvoiced)
Sibilants S, SH, Z, ZH and affricates CH, JH
Nasals
Semivowels
Aspirate
Sonorants
Regular formant structure vowels and L, R, M, N
Noncontinuants
Diphthongs
Stops
Affricates
Obstruents
Fricatives (voiced and unvoiced)
Stops

A.2 Phoneme Classification with Phonetic Alphabet and Key Words

AMERICAN ENGLISH PHONEMES



Vowels

Back

UW /u/ boot
UH /ʊ/ book
OW /o/ boat
AO /ɔ/ bought

Mid

AA /ɑ/ Bach
ER /ɝ/ Burt
AH /ʌ/ but

Front

IY /i/ beet
IH /ɪ/ bit
EH /ɛ/ bet
AE /æ/ bat

Additional Vowels

EY /e/ bait
AX /ə/ alone
IX /ɨ/ debit

Diphthongs

AY /aɪ/ buy
OY /ɔɪ/ boy
AW /aʊ/ down

Consonants

Semivowels

Liquids

R /r/ rye
L /l/ lie

Glides

W /w/ why
Y /j/ yes

Nasals

M /m/ my
N /n/ night
NX /ŋ/ rang

Affricates

CH /tʃ/ chair
JH /dʒ/ judge

Aspirate

HH /h/ hat

Plosives

Voiced

B /b/ by
D /d/ dye
G /g/ guy

Unvoiced

P /p/ pie
T /t/ tie
K /k/ key

Fricatives

Voiced

V /v/ van
DH /ð/ this
Z /z/ zoo
ZH /ʒ/ azure

Unvoiced

F /f/ fix
TH /θ/ thick
S /s/ sat
SH /ʃ/ ship

Note: There is not universal agreement concerning the classification of vowels and diphthongs. In addition some authors include additional allophones of various phonemes.

APPENDIX B
ARTICULATORY SPEECH CODING

B.1 Speech Coding Algorithm

A new approach to very–low–rate speech coding can be proposed that will focus on the fact that the articulators vary slowly with time, and, therefore, can be utilized to segment speech for the purpose of speech coding. The proposed technique is to segment speech according to phonetically related intervals that are synchronized with articulatory movement and, thus, to establish a relationship between speech segmentation and articulatory movement. If the segments are longer than those presently used with the pitch asynchronous method, then we may be able to code speech according to articulatory movement and achieve a low bandwidth, high quality speech coding scheme. The articulatory position and its movement can be estimated using the acoustic information obtained by glottal inverse filtering and an acoustic–to–articulatory transformation technique.

As speech is generated the articulators change their position slowly compared with the rate of vibration of the vocal folds. If the position of the articulators (i.e. the vocal tract cross sectional area) can be estimated from the speech signal, then it can be used as a criteria for speech segmentation. One of the difficulties is the problem of estimating the articulatory parameters from the speech signal.

Recently, this topic has become of interest to researchers because the quality of known coding algorithms at low bit rates is unacceptably poor and the computing power required to implement the complex algorithms based on the articulatory production model is now available.

We have developed a method for calculating the position of the articulators that minimizes the error between the formants measured from the speech signal by the analysis tool and the formants calculated from the articulatory vocal tract configuration (Prado, 1991; Hsieh, 1994). In other words, once we analyze the speech signal to get the formant information, the position of the articulators can be estimated by the speech inverse filtering procedure.

The new speech coding algorithm based on the movement of the articulators is summarized in Figure B–1. Assuming we know the previous estimates for the articulator positions, we predict the new positions for the next analysis frame based upon their present and past positions and their velocity of movement. Here, the present position of the articulator can be estimated by applying the speech inverse filtering technique to the formant information obtained from speech analysis. The speech analysis part will be explained later. Estimates of the formants for the next analysis frame can be predicted from the vocal tract area function, which is calculated from the predicted articulatory position. The error between these predicted formants and new formants estimated from the speech signal is one of our criteria for speech segmentation. When the error is less than a threshold value, then the articulator presumably did not move greatly in the speech interval compared to the previous one. When the error is larger than the threshold value, we update the articulatory vector, segment it, and store the vectors. This could prove to be a very low bit rate speech coding scheme (probably less than 2400 bits per second).

B.2 Speech Reconstruction (Decoding) Algorithm

The original speech can be reconstructed from the glottal source parameters and articulatory position vectors. Figure B–2 shows a block diagram of the speech reconstruction procedure using the formant synthesizer. The glottal source waveform input is recovered from the LF model parameters and the jitter and shimmer model

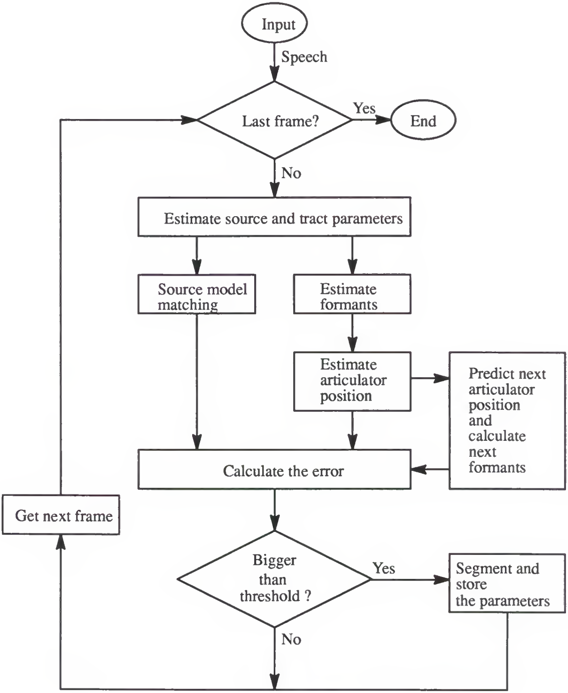

Input
Speech
Last frame?
Yes
End
No
Estimate source and tract parameters
Source model matching
Estimate formants
Estimate articulator position
Predict next articulator position and calculate next formants
Calculate the error
Get next frame
Bigger than threshold ?
Yes
Segment and store the parameters
No

Figure B−1. An articulatory speech coding algorithm.

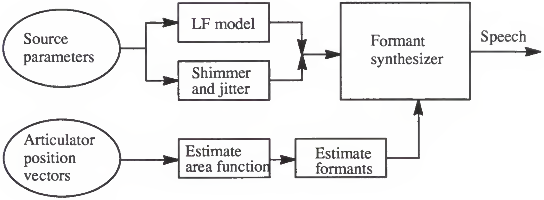


Figure B–2. Speech reconstruction using a formant synthesizer.

parameters. The vocal tract area function is calculated from the articulator position vector and the formant frequencies and bandwidths can be solved for using the area function.

Figure B–3 shows another possible speech reconstruction scheme in which the articulatory synthesizer is used. Since the articulatory synthesizer only requires the articulatory position and glottal source as input, the configuration scheme is quite simple.

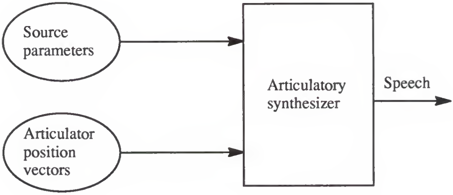


Figure B–3. Speech reconstruction using an articulatory synthesizer.

REFERENCES

Acronyms:

ASSP – Acoustics, Speech, and Signal Processing
IEEE – Institute of Electrical and Electronic Engineers
STL–QPSR, RIT – Speech Transmission Lab. Quarterly Progress and Status Report, Royal Inst. of Tech.

Alku, P. (1992). "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering," Speech Communication, 11, 109–118.

Allen, D. R., and Strong, W. J. (1985). "A model for the synthesis of natural sounding vowels," J. Acoust. Soc. Am., 78(1), 58–69.

Atal, B. S., Chang, J. J., Mathews, M. V., and Tukey, J. W. (1978). "Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique," J. Acoust. Soc. Am., 63(5), 1535–1555.

Atal, B. S., and Hanauer, S. L. (1971). "Speech analysis and synthesis by Linear Prediction of the speech wave," J. Acoust. Soc. Am, 50(2), 637–655.

Atal, B. S., and Remde, J. R. (1982). "A new model of LPC excitation for producing natural-sounding speech at low bit rates," Proc. IEEE Int. Conf. on ASSP, 614–617.

Badin, P. (1989). "Acoustics of voiceless fricatives: Production theory and data," STL–QPSR, RIT, Stockholm, Sweden, 3, 33–55.

Badin, P. (1991). "Fricative consonants: Acoustic and X-ray measurements," J. of Phonetics, 19, 397–408.

Badin, P., and Fant, G. (1984). "Notes on vocal tract computation," STL–QPSR, RIT, Stockholm, Sweden, 2–3, 53–108.

Bladon, R. A. W., and Fant, G. (1978). "A two–formant model and cardinal vowels," STL–QPRS, RIT, Stockholm, Sweden, 1, 1–8.

Bocchieri, E. L., and Childers, D. G. (1984). "Interactive graphics editor permits study of animated speech articulation," Speech Technology, 2(2), 10–14.

Broad, D. J. (1977). "Theory of vocal fold vibration," in Topics in Speech Science, edited by D. J. Broad (Speech Communications Research Laboratory, Inc., Los Angeles, CA), 158–232.

Carlson, R., Fant, G., and Granstrom, B. (1975). "Two–formant models, pitch and vowel perception," Auditory Analysis and Perception of Speech, Academic Press, London, 55–82.

Charpentier, F. (1984). "Determination of the vocal tract shape from the formants by analysis of the articulatory-to-acoustic nonlinearities," Speech Communication, 3, 291–308.

Childers, D. G. (1991). "Signal processing methods for the assessment of vocal disorders," Medical and Life Sciences Engineering, 13, 117–130.

Childers, D. G., and Ahn, C (1995). "Modeling the glottal volume–velocity waveform for three voice types," J. Acoust. Soc. Am, 97(1), 505–519.

Childers, D. G., and Hu, H. T. (1994). "Speech synthesis by glottal excited linear prediction," J. Acoust. Soc. Am., 96(4), 2026–2036.

Childers, D. G. and Krishnamurthy, A. K. (1985). "A critical review of electroglottography," CRC Critical Reviews in Biomedical Engineering, 12(2), 131–164.

Childers, D. G., and Lee, C. K. (1991). "Vocal quality factors: Analysis, synthesis, and perception," J. Acoust. Soc. Am, 90(5), 2394–2410.

Childers, D. G., and Wu, K. (1990). "Quality of speech produced by analysis–synthesis," Speech Communication, 9, 97–117.

Coker, C. H. (1976). "A model of articulatory dynamics and control," Proc. IEEE, 64(4), 452–460.

Delattre, P., Liberman, A. M., Cooper, F. S., and Gerstman, L. J. (1952). "An experimental study of the acoustic determinants of vowel color," Word, 8, 195–210.

Eskenazi, L., Childers, D. G., and Hicks, D. M. (1990). "Acoustic correlates of vocal quality," J. Speech and Hearing Research, 33, 298–306.

Fant, G. (1960). Acoustic Theory of Speech Production, Mouton, The Hague.

Fant, G. (1978). "Vowel perception and specification," Rivista Italiana di Acustica, II, N. 2, 69–84.

Fant, G. (1979). "Glottal source and excitation analysis," STL–QPSR, RIT, Stockholm, Sweden, 1, 85–107.

Fant, G. (1985). "The vocal tract in your pocket calculator," STL–QPSR, RIT, Stockholm, Sweden, 2–3, 1–19.

Fant, G. (1986). "Glottal flow: Models and interaction," J. Phonetics, 14, 393–400.

Fant, G., Liljencrants, J., and Lin, Q. G. (1985). "A four–parameter model of glottal flow," STL–QPSR, RIT, Stockholm, Sweden, 4, 1–13.

Fant, G., and Lin, Q. G. (1987). "Glottal source – vocal tract acoustic interaction," STL–QPSR, RIT, Stockholm, Sweden, 1, 13–27.

Fant, G., and Risberg, A. (1962). "Auditory matching of vowels with two formant synthetic sounds," STL–QPSR, RIT, Stockholm, Sweden, 4, 7–11, .

Fant, G., and Risberg, A. (1963). "Auditory matching of vowels with two formant synthetic sounds," STL–QPSR, RIT, Stockholm, Sweden, 4, 7–11.

Fettweis, A., and Meerkötter, K. (1975). "On adaptors for wave digital filters," IEEE Trans. on ASSP, 23(6), 516–525.

Flanagan, J. L., and Cherry, L. (1968). "Excitation of vocal-tract synthesizers," J. Acoust. Soc. Am., 45(3), 764–769.

Flanagan, J. L., and Ishizaka, K. L. (1976). "Automatic generation of voiceless excitation to a vocal cord–vocal tract speech synthesizer," IEEE Trans. on ASSP, 24(2), 163–170.

Flanagan, J. L., Ishizaka, K., and Shipley, K. L. (1975). "Synthesis of speech from a dynamic model of the vocal cords and vocal tract," Bell System Technical Journal, 54, 485–506.

Flanagan, J. L., and Landgraf, I. L. (1968). "Self-oscillating source for vocal tract synthesizers," IEEE Trans. on Audio and Electroacoustics, 16, 57–64.

Flanagan, J. L., Ishizaka, K., and Shipley, K. L. (1980). "Signal models for low bit–rate coding of speech," J. Acoust. Soc. Am., 68(3), 780–791.

Fujisaki, H., and Ljungqvist, M. (1986). "Proposal and evaluation of models for the glottal source waveform," Proc. IEEE Int. Conf. on ASSP, 1605–1608.

Fujisaki, H., and Ljungqvist, M. (1987). "Estimation of voice source and vocal tract parameters based on ARMA and a model for the glottal source waveform," Proc. IEEE Int. Conf. on ASSP, 637–640.

Ghitza, O. (1994). "Auditory models and human performance in tasks related to speech coding and speech recognition," IEEE Trans. on Speech and Audio Processing, 2(1), 115–132.

Gobl, C. (1988). "Voice source dynamics in connected speech," STL–QPSR, RIT, Stockholm, Sweden, 1, 123–159.

Heinz, J. M., and Stevens, K. N. (1961). "On the properties of voiceless fricatives," J. Acoust. Soc. Am., 33, 589–596.

Hermansky, H. (1986). "Perceptually based processing in automatic speech recognition," Proc. IEEE Int. Conf. on ASSP, 1971–1974.

Hermansky, H. (1987). "An efficient speaker–independent automatic speech recognition by simulation of some properties of human auditory perception," Proc. IEEE Int. Conf. on ASSP, 1159–1162.

Hermansky, H. (1990). "Perceptual linear predictive (PLP) analysis of speech," J. Acoust. Soc. Am., 87(4), 1738–1752.

Hermansky, H., and Junqua, J. C. (1988). "Optimization of perceptually based ASR front end," Proc. IEEE Int. Conf. on ASSP, 219–222.

Hermansky, H., and Broad, D. J. (1989). "The effective second formant F2' and the vocal tract front–cavity," Proc. IEEE Int. Conf. on ASSP, 480–483.

Hsieh, Y. F. (1994). "A flexible and high quality articulatory speech synthesizer," Ph.D. Dissertation, University of Florida.

Hu, H. T. (1993). "An improved source model for a linear prediction speech synthesizer," Ph.D. Dissertation, University of Florida.

Johansson, C., Sundberg, J., Wilbrand, H., and Ytterbergh, C. (1983). "From sagittal distance to area," STL–QPSR, RIT, Stockholm, 4, 39–49.

Kay, S. M. (1988). Modern Spectral Estimation, Prentice–Hall, Englewood Cliffs, NJ.

Kelly, J. K., Jr., and Lochbaum, C. C. (1962). "Speech synthesis," Proc. Fourth Intern. Congr. Acoust., G42, 1–4.

Kewley–Port, D. (1982). "Measurement of formant transitions in naturally produced stop consonant–vowel syllables," J. Acoust. Soc. Am., 72(2), 379–389.

Kewley–Port, D. (1983). "Time–varying features as correlates of place of articulation in stop consonants," J. Acoust. Soc. Am., 73(1), 322–335.

Kewley–Port, D., and Luce, P. (1984). "Time–varing features of initial stop consonants in a auditory running spectra: A first report," Percept. Psychophys., 35, 353–360.

Klatt, D. H. (1980). "Software for a cascade/parallel formant synthesizer," J. Acoust. Soc. Am., 67(3), 971–995.

Klatt, D. H. (1987). "Review of text–to–speech conversion for english," J. Acoust. Soc. Am., 82(3), 737–793.

Klatt, D. H., and Klatt, L. C. (1990). "Analysis, synthesis, and perception of voice quality variations among female and male talkers," J. Acoust. Soc. Am., 87(2), 820–857.

Kuhn, G.M. (1975). "On the front cavity resonance and its possible role in speech perception," J. Acoust. Soc. Am., 58, No. 2, 428–433.

Kurowski, K., and Blumstein, S. (1984). "Perceptual integration of the murmur and formant transitions for place of articulation in diffuse nasal consonants," J. Acoust. Soc. Am., 76(2), 383–390.

Kurowski, K., and Blumstein, S. (1987). "Acoustic properties for place of articulation in nasal consonants," J. Acoust. Soc. Am., 81(6), 1917–1927.

Lawson, S., and Mirzai, A. R. (1990). Wave Digital Filters, E. Horwood, New York.

Lee, C. K. (1988). "Voice quality: Analysis and synthesis," Ph.D. Dissertation, University of Florida.

Lee, K. S. (1992). "Pitch synchronous analysis/synthesis using the WRLS–VFF–VT algorithm," Ph.D. Dissertation, University of Florida.

Levinson, S. E., and Schmidt, C. E. (1983). "Adaptive computation of articulatory parameters from the speech signal," J. Acoust. Soc. Am., 74(4), 1145–1154.

Lin, Q. G. (1990). "Speech production theory and articulatory speech synthesis," Ph.D. dissertation, RIT, Stockholm, Sweden.

Lin, Q. G. (1992). "Vocal-tract computation: How to make it more robust and faster," STL–QPSR, RIT, Stockholm, Sweden, 4, 29–42.

Lobo, A. P., and Ainsworth, W. A. (1992). "Evaluation of a glottal ARMA model of speech production," Proc. IEEE Int. Conf. on ASSP, II, 13–16.

Markel, J. D., and Gray, A. H. (1976). Linear Prediction of Speech, Springer–Verlag, New York.

Mermelstein, P. (1973). "Articulatory model for the study of speech production," J. Acoust. Soc. Am., 53(4), 1070–1082.

Meyer, P., Wilhelms, R., and Strube, H. W. (1989). "A quasiarticulatory speech synthesizer for German language running in real time," J. Acoust. Soc. Am., 86(2), 523–539.

Milenkovic, P .H. (1986). "Glottal inverse filtering by joint estimation of an AR system with a linear input model," IEEE Trans. Acoust., Speech, and Signal Processing, 34(1), 28–42.

Milenkovic, P .H. (1987). "Acoustic tube reconstruction from noncausal excitation," IEEE Trans. Acoust., Speech, and Signal Processing, 35(8), 1089–1100.

Milenkovic, P. H. (1993). "Voice source model for continuous control of pitch period," J. Acoust. Soc. Am., 93(2), 1087–1096.

Oppenheim, A. V., and Willsky, A. S. (1983). Signals and Systems, Prentice–Hall, Englewood Cliffs, NJ.

Parthasarathy, S., and Coker, C. H. (1990). "Phoneme-level parameterization of speech using an articulatory model," Proc. IEEE Int. Conf. on ASSP, 337–340.

Parthasarathy, S., and Coker, C. H. (1992). "On automatic estimation of articulatory parameters in a text-to-speech system," Computer Speech and Language, 6, 37–75.

Prado, P. P. L. (1991). "A target-based articulatory synthesizer," Ph.D. dissertation, University of Florida.

Qi, Y., and Fox, R. A. (1992). "Analysis of nasal consonants using perceptual linear prediction," J. Acoust. Soc. Am, 91(3), 1718–1726.

Rabiner, L. R., and Schafer, R. W. (1978). Digital Processing of Speech Signals, Prentice–Hall, Englewood Cliffs, NJ.

Rothenberg, M. R., Calson, R., Granstrom, B., and Gaufin, J. (1975). "A three–parameter voice source for speech synthesis," Speech Communication, 2, 235–243.

Rothenberg, M. R. (1981). "An interactive model for the voice source," STL–QPSR, RIT, Stockholm, Sweden, 4, 1–17.

Rye, J. M. and Holmes, J. N. (1982). "A versatile software parallel formant speech synthesizer," JSRU Research Report, No. 1016.

Schroeder, M. (1977). "Recognition of complex acoustic signals," Life Science Research Report 5, edited by T. Bullock (abakon Verlag, Berlin), 324.

Schroeter, J., and Sondhi, M. M. (1994). "Techniques for estimating vocal–tract shapes from the speech signal," IEEE Trans. on Speech and Audio Processing, 2(1), 133–150.

Singhal, S., and Atal, B. S. (1989). "Amplitude optimization and pitch prediction in multipulse coders," IEEE Trans. on ASSP, 37(3), 317–327.

Soli, S. D. (1981). "Second formants in fricatives: Acoustic consequences of fricative –vowel coarticulation," J. Acoust. Soc. Am., 70(4), 976–984.

Sondhi, M. M. (1975). "Measurement of the glottal waveform," J. Acoust. Soc. Am, 57(1), 228–232.

Sondhi, M. M., and Resnick, J. R. (1983). "The inverse problem for the vocal tract: Numerical methods, acoustical experiments, and speech synthesis," J. Acoust. Soc. Am., 73(3), 985–1022.

Sondhi, M. M., and Schroeter, J. (1986). "A nonlinear articulatory speech synthesizer using both time- and frequency-domain elements," Proc. IEEE Int. Conf. on ASSP, 1999–2002.

Sondhi, M. M., and Schroeter, J. (1987). "A hybrid time-frequency domain articulatory speech synthesizer," IEEE Trans. on ASSP, 35(7), 955–967.

Sorokin, V. N. (1985). Theory of Speech Formation, Teoriya Recheobrazovaniya, Publishing House "Radio i Svyaz'", Moscow.

Sorokin, V. N. (1992). "Determination of vocal tract shape for vowels," Speech Communication, 11, 71–85.

Sorokin, V. N. (1994). "Inverse problem for fricatives," Speech Communication, 14, 249–262.

Stevens, K. N. (1971). "Airflow and turbulence noise for fricative and stop consonants: Static considerations," J. Acoust. Soc. Am, 50(4), 1180–1192.

Stevens, K. N. (1993a). "Modelling affricate consonants," Speech Communication, 13, 33–43.

Stevens, K. N. (1993b). "Models for the production and acoustics of stop consonants," Speech Communication, 13, 367–375.

Stevens, S. S. (1957). "On the psychophysical law," Psychol. Rev., 64, 153–181.

Ting, Y. T. (1989). "Adaptive estimation of time–varying signal parameters with applications to speech," Ph.D. Dissertation, University of Florida.

Yu, Z. (1993). "A method to determine the area function of speech based on perturbation theory," STL–QPSR, RIT, Stockholm, Sweden, 4, 77–95.

Veeneman, D. E., and B$_E$Ment, S. L. (1985). "Automatic glottal inverse filtering from speech and electroglottographic signals," IEEE Trans. on ASSP, 33(2), 369–377.

Wakita, H., and Fant, G. (1978). "Toward a better vocal tract model," STL–QPSR, RIT, Stockholm, Sweden, 1, 9–29.

Wong, D. Y., Markel, J. D., and Gray, A. H. Jr. (1979). "Least squares glottal inverse filtering from the acoustic speech waveform," IEEE Trans. on ASSP, 27(4), 350–355.

Xue, Q., Hu, Y. H., and Milenkovic, P. (1990). "Analyses of the hidden units of the multi-layer perceptron and its application in acoustic-to-articulatory mapping," Proc. IEEE Int. Conf. on ASSP, 869–872.

Zwicker, E., and Fastl, H. (1990). Psychoacoustics: Facts and models, Springer–Verlag, Berlin.

Zwicker, E., and Zwicker, U. T. (1991). "Audio engineering and psychoacoustics: Matching signals to the final receiver, the human auditory system," J. Audio Engineering Soc., 39(3), 115–126.

BIOGRAPHICAL SKETCH

Mr. Minkyu Lee was born in Korea on October 6, 1963. He graduated from the Seoul National University, Seoul, Korea, in February, 1986, with a B.Sc. degree in electrical engineering. He also received his M.S. degree from the Korea Advanced Institute of Science and Technology (KAIST), Seoul, Korea, in February, 1988. He joined the Electronics and Telecommunications Research Institute (ETRI), Taejon, Korea, as a member of the research staff in March, 1988. He had been involved in research and development of the Message Handling System based on X.400 CCITT recommendation.

Since August, 1992, he has been with the Department of Electrical and Computer Engineering, University of Florida, Gainesville, Florida, where his primary areas of interest are digital signal processing, speech analysis and synthesis, speech coding, speech recognition and computer engineering.

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Donald G. Childers

Donald G. Childers, Chairman
Professor of Electrical and Computer Engineering

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Frederick J. Taylor
Professor of Electrical and Computer Engineering

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Jose C. Principe
Professor of Electrical and Computer Engineering

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

L. W. Couch

Leon W. Couch
Professor of Electrical and Computer Engineering

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Howard B. Rothman

Howard B. Rothman
Professor of Communication Processes
and Disorders

This dissertation was submitted to the Graduate Faculty of the College of Engineering and to the Graduate School and was accepted as partial fulfillment of the requirements for the degree of Doctor of Philosophy.

May, 1996

for Winfred M. Phillips
Dean, College of Engineering

Karen A. Holbrook
Dean, Graduate School